DNASpeech: A Contextualized and Situated Text-to-Speech Dataset with Dialogues, Narratives and Actions

Anonymous ACL submission

Abstract

In this paper, we propose contextualized and situated text-to-speech (CS-TTS), a novel TTS task to promote more accurate and customized speech generation using prompts with **D**ialogues, **N**arratives, and **A**ctions (DNA). While prompt-based TTS methods facilitate controllable speech generation, existing TTS datasets lack situated descriptive prompts aligned with speech data. To address this data scarcity, we develop an automatic annotation pipeline enabling multifaceted alignment among speech clips, content text, and their respective descriptions. Based on this pipeline, we present DNASpeech, a novel CS-TTS dataset with high-quality speeches with DNA prompt annotations. DNASpeech contains **2,395 distinct characters, 4,452 scenes, and 22,975 dialogue utterances**, along with over **18 hours of high-quality speech recordings**. To accommodate more specific task scenarios, we establish a leaderboard featuring two new subtasks for evaluation: CS-TTS with narratives and CS-TTS with dialogues. We also design an intuitive baseline model for comparison with existing state-of-the-art TTS methods on our leaderboard. Experimental results indicate the quality and effectiveness of DNASpeech, validating its potential to drive advancements in the TTS field. Dataset is available at `https://anonymous.4open.science/r/DNASpeech-FDCD` [1]

1 Introduction

Text-to-speech (TTS) aims to convert input text into human-like speech, attracting significant attention in the audio and speech processing community (Shen et al., 2018; Ren et al., 2020; Shen et al., 2023; Ju et al., 2024). Previous studies have shown that incorporating more detailed descriptions of the input text is crucial for improving the accuracy of speech synthesis (Guo et al., 2023; Li et al., 2022b; Yang et al., 2024). The speaker's contextual information, such as dialogue history, significantly impacts the generated speech (Li et al., 2022a; Guo et al., 2021; Liu et al., 2023). Additionally, situated descriptions are also beneficial to enhance the expressiveness of the speech by providing environmental background (Lee et al., 2024). Consequently, we propose a new TTS task termed Contextualized and situated Text-To-Speech (CS-TTS), which considers the impact of contextualized and situated descriptions on speech synthesis. By integrating these detailed descriptions, CS-TTS enables more accurate and expressive speech generation, improving the applicability of TTS systems across diverse scenarios.

Recently, prompt-based TTS methods have gained increasing research interest, providing technical support for customized speech generation (Li et al., 2024). While formulating detailed descriptions as prompts can potentially address the CS-TTS task, current datasets lack comprehensive prompts that align with text and speech. Their limitations include: (1) Existing prompts with several key phrases lack sufficient contextual descriptions (Kim et al., 2021; Guo et al., 2023); (2) Dialogue-only prompts fail to incorporate multifaceted situated descriptions required for precise speech customization (Lee et al., 2023; Li et al., 2022a); (3) Limited speaker characters restrict the exploration of various acoustic characteristics in TTS generation.

These constraints render existing datasets insufficient for CS-TTS research. Therefore, we aim to construct a new CS-TTS dataset incorporating more comprehensive contextualized and situated descriptions. As illustrated in Figure 1, we systematically summarize the necessary descriptions into three categories, abbreviated as "**DNA**": **Dialogues** provide the conversational context of speech content; **Narratives** describe the environmental scenes surrounding the speaker's speech; and **Actions** de-

[1] Dataset will be made public once accepted.

Figure 1: An illustration of **DNASpeech Dataset**. "DNA" descriptions for our proposed CS-TTS task. Dialogues, Narratives, and Actions are annotated to capture the contextualized and situated background essential for TTS generation.

tail the speaker's actions and expressions during speech production.

Among various data sources, movies offer a natural solution due to their rich speech content and diverse character timbres. Movie scripts include not only conversational lines but also environmental scenes that guide the speaker's performance, aligning well with our "DNA" descriptions. Taking advantage of this, we develop an automated annotation pipeline for multifaceted alignment among content text, speech clips, and their corresponding "DNA" descriptions. Based on our efforts in processing movie videos and scripts through this pipeline, we finally collect a new CS-TTS dataset DNASpeech that contains 2,395 distinct characters, 4,452 scenes, and 22,975 dialogue utterances, along with over 18 hours of high-quality speech recordings.

To accommodate more specific task scenarios, we establish a leaderboard featuring two new subtasks: CS-TTS with narratives and CS-TTS with dialogues. Both subtasks are used to evaluate the ability of TTS systems to leverage environmental scenes and dialogue context, along with the speaker's actions, to customize speech. We also introduce an intuitive CS-TTS baseline model for comparison with existing representative TTS methods on our leaderboard. Extensive experimental results validate the effectiveness and quality of DNASpeech, contributing to the advancements of prompt-based TTS.

Our main conclusions can be summarized as follows:

(1) To support research in CS-TTS, we collect a novel dataset DNASpeech, containing high-quality speech recordings annotated with comprehensive "DNA" prompts: dialogues, narratives, and actions.

(2) We elaborately present an automatic annotation pipeline for multifaceted alignment among content text, speech clips, and their corresponding descriptions, enabling the efficient collection of high-quality aligned TTS data.

(3) We establish a leaderboard featuring two new subtasks: CS-TTS with narratives and CS-TTS with dialogues. We also propose an intuitive baseline model for the CS-TTS task. Comprehensive experimental results indicate the quality and effectiveness of DNASpeech.

2 Related Work

2.1 Text-to-speech without prompts

Text-to-speech (TTS) systems have been significantly propelled by the availability of diverse and extensive speech datasets. LJSpeech (Ito and Johnson, 2017) stands out with its 13,100 high-quality short speech clips of a single speaker, derived from readings of passages from seven non-fiction books. Another key resource is the LibriSpeech corpus (Panayotov et al., 2015), an extensive collection encompassing approximately 1,000 hours of audiobook recordings from the LibriVox project (Kearns, 2014).

To expand these resources, LibriTTS (Zen et al., 2019) offers a multi-speaker English corpus with around 585 hours of read speech, recorded at a 24kHz sampling rate, enhancing the variability and richness of the speech data available for TTS research. The CSTR VCTK Corpus [2] further diversifies the available data with contributions from 110 English speakers exhibiting various accents, each providing approximately 400 sentences sourced from diverse texts, such as newspapers and accent elicitation passages. Moreover, the Hi-Fi Multi-Speaker English TTS Dataset (Hi-Fi TTS) (Bakhturina et al., 2021) delivers a robust multi-speaker dataset, consisting of approximately 291.6 hours of speech from 10 speakers, with each contributing at least 17 hours of recordings. These datasets collectively furnish a rich foundation for developing and refining TTS systems, enabling significant improvements in the naturalness and intelligibility of synthetic speech.

2.2 Text-to-speech with prompts

With the advancement of TTS technology, there has been an increasing emphasis on using prompts to guide speech generation, enabling a more diverse and customized generation process. Initially, seminal works (Adigwe et al., 2018; Livingstone and Russo, 2018; Zhou et al., 2021) identify the presence of emotional information in speech and construct corresponding datasets by annotating speech with emotions. However, these datasets primarily focus on emotional labels within speech and categorize them into a limited number of classes. To achieve more comprehensive representations, FSNR0 (Kim et al., 2021) introduces 327 different labels covering a variety of emotions, intentions, tones, and speech rates. To further advance prompt-based TTS, the PromptSpeech dataset from PromptTTS (Guo et al., 2023) utilizes continuous text to describe speech across multiple dimensions, including gender, pitch, loudness, speech rate, and emotion. Similarly, NLSpeech (Yang et al., 2024) and TextrolSpeech (Ji et al., 2024) employ continuous text descriptions of speech, incorporating more detailed and daily expressions.

The datasets mentioned above mainly focus on describing the speech, lacking contextual information crucial for speech generation. Despite these advancements, datasets with contextual prompts remain relatively scarce. DailyTalk (Lee et al., 2023) is a highly popular dataset consisting of 20 hours of speech data from 2,541 dialogues, spoken by two fluent English speakers, a male and a female. The dialogues in DailyTalk are sampled from another dialogue dataset DailyDialog (Li et al., 2017). ECC (Li et al., 2022a) collects 24 hours of speeches from 66 conversational videos from YouTube. Each dialogue has a duration of 79.3 seconds and features around 2.9 speakers on average. In contrast, MM-TTS (Li et al., 2024) highlights the influence of environmental information on speech, amassing expressive speech from film and television data, aligned with corresponding facial expressions and actions.

Unlike existing contextual prompt-based TTS datasets (Lee et al., 2023; Li et al., 2022a, 2024), our DNASpeech systematically integrates and aligns three distinct types of descriptive prompts, providing more comprehensive contextualized and situated information to enhance the richness and relevance of the generated speech. Moreover, DNASpeech presents a substantial enhancement in speaker diversity, enabling the exploration of various acoustic characteristics in TTS generation.

3 DNASpeech Dataset

3.1 Overview

What is DNASpeech? We aim to construct a pioneering prompt-based TTS dataset tailored for the CS-TTS task. The proposed dataset DNASpeech aggregates a significant corpus of speech clips

[2] https://datashare.ed.ac.uk/handle/10283/3443

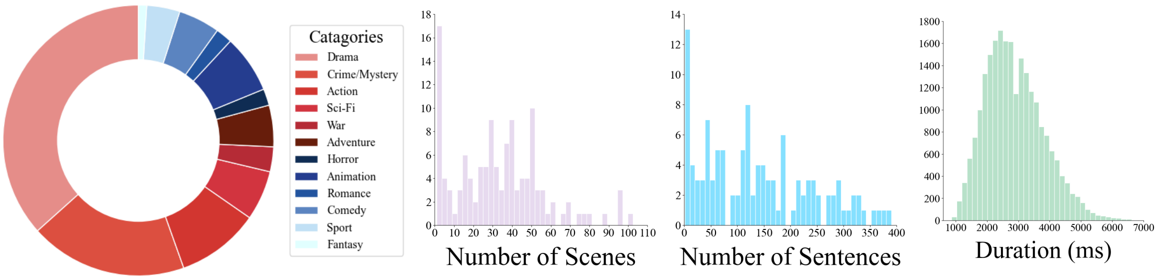


Figure 2: **The DNASpeech Dataset.** *Pie Chart:* Proportion of movie categories. *Histograms, from left to right:* Distribution of the number of scenes, sentences, and speech clip duration in movies. Best viewed online and zoomed in.

sourced from movies and their accompanying scripts. Each speech clip is aligned with three types of prompts: dialogues (D), narratives (N), and actions (A). These prompts, collectively referred to as "DNA", are intricately intertwined with the corresponding speeches, enhancing the contextual richness and situational relevance of the dataset. Specifically, dialogues contain the conversational context preceding the speech; narratives depict the environmental scenes surrounding the speech; and actions describe the speaker's actions and expressions during speech production.

Why are contextualized and situational prompts necessary? Textual prompts serve as crucial directives for controlling speech generation, guiding the extraction of emotional and acoustic features necessary for speech synthesis. However, current datasets typically employ direct prompts, which explicitly describe the desired speech attributes such as "Angry, High pitch, Low speed, Loudly." These prompts essentially function as speech annotations and may not always be readily available, particularly in scenarios like audiobooks where detailed prompts are lacking (Anguera et al., 2011). In contrast, contextual prompts are closely associated with speech and reflect the situational context in which the speech occurs. For instance, the speech in a spooky and fearful scene is expected to convey low-pitched and tense tones. Despite their prevalence, datasets incorporating such contextualized and situated prompts remain scarce in the field of TTS. Moreover, contextualized prompts require TTS systems to identify subtle nuances of the surrounding context. Therefore, the inclusion of contextual prompts holds promise for driving advancements in TTS by enabling more contextually appropriate and natural speech synthesis.

3.2 Dataset Construction Pipeline

To efficiently and automatically annotate descriptive prompts aligned with text and speech, we develop a new annotation pipeline. Fig 3 illustrates the overview of this pipeline for DNASpeech, which consists of five fundamental steps: (1) data collection, (2) information extraction, (3) cross-modal alignment, (4) speech denoising, and (5) automatic speech recognition. Data collection and information extraction provide and preprocess the raw movie materials. Cross-modal alignment integrates speech and textual descriptions through both coarse-grained and fine-grained alignment processes. Speech denoising and automatic speech recognition ensure the quality of the speeches.

Step 1: Data Collection Movies serve as an invaluable resource for TTS research due to their rich speech data and detailed contextual information found in corresponding scripts, such as dialogue lines, narrative scenes, and action depictions. Therefore, we choose movies as the primary data source to construct DNASpeech.

Inspired by the Condensed Movies Dataset (CMD) (Bain et al., 2020) compiling a substantial collection of licensed movie clips from the MovieClip YouTube channel [3], we augment our dataset by collecting newly uploaded movies from the MovieClip channel and purchasing additional movies from legitimate sources. Eventually, we collect a total of 126 movies released between 1940 and 2023, spanning up to 14 common movie categories, to enrich the diversity of our dataset.

Step 2: Information Extraction Following collecting the raw movie videos, the next step is to extract the necessary information, including the

[3] https://www.youtube.com/c/MOVIECLIPS

speaker's voice and its corresponding lines. Subtitles in SRT format [4] contain the content text along with timestamps for the start and end of each speech segment. We leverage timestamps to obtain aligned text-speech pairs. For other subtitles in image format, we employ SubtitleEdit[5], a widely used software to convert image subtitles into text format using Optical Character Recognition (OCR) technology. Once all subtitles are converted into SRT format, we extract the corresponding speech clips from the movie soundtracks, sampled at a rate of 16,000 Hz, thus obtaining both the speech clips and their associated content text.

Next, our focus shifts to movie scripts obtained from the Internet Movie Script Database (IMSDb)[6], a comprehensive repository of thousands of movie scripts. However, original movie scripts are lengthy and unstructured, necessitating parsing into structured units. Following the script writing paradigm, we extract four key elements from each movie script: *Dialogues Narratives*, *Actions*, and *Characters*. Dialogues denote the speaker's conversational context and line content of their speech within a scene. Narratives represent the basic units defining the overall setting of a shot in the movie. Actions provide supplementary details about characters, describing their actions and expressions. Characters denote the actors for each conversational session. This process allows us to gather the contextualized and situated information of speeches in movies.

Step 3: Cross-modal Alignment Prompt-based TTS tasks necessitate aligning each speech with its corresponding prompts, which is crucial for effective speech synthesis. Leveraging the shared content text between speeches and lines provides a foundation for tackling this alignment challenge. However, while it is theoretically straightforward, aligning speeches with lines directly from the script encounters discrepancies in the content text. To address this issue, we implement a two-stage alignment module combining coarse-grained and fine-grained alignment.

coarse-grained alignment. To match each speech with its corresponding line in the script, more than 800 million potential matches are required, which is computationally intensive and increases the cost of manual verification. Hence, we initially filter out pairs with low textual similarity by performing coarse-grained matching. To be more specific, we preprocess both speech and script content by removing stop words, punctuation, and lemmatizing words. We then employ the Longest Common Subsequence (LCS) method to compute textual similarity, retaining *(speech, text)* pairs with a similarity score of 0.9 or higher for subsequent fine-grained alignment.

fine-grained alignment. After coarse-grained alignment, we obtain approximately 30,000 *(speech, text)* pairs. However, the overlap between textual strings may not adequately capture the alignment degree between speech and text. Therefore, in this stage, we utilize the official sentence model `all-mpnet-base-v2`[7] presented by sentence-transformers group to calculate the semantic similarity between speech and text. Pairs with a semantic similarity score of 0.7 or higher are retained. Finally, this process yields 22,975 *(speech, text)* pairs, totaling 18.37 hours of speech data.

Step 4: Speech Denoising The speech clips extracted from the movies in Step 2 usually contain background noises that degrade the quality of the human voice. Therefore, it is essential to separate the human voice from the background noise. Additionally, the speech may sometimes be unclear due to the filming environment, which makes it also important to further enhance the human voice. To eliminate these disturbing noises, we employed Resemble Enhance[8], a common tool designed for noise reduction and speech enhancement. This tool comprises a denoiser and an enhancer, which extract human voices from complex background noise and further improve perceived audio quality by restoring audio distortions and extending the audio bandwidth. Both models are trained using high-quality 44.1kHz voice data, ensuring superior speech enhancement.

Step 5: Automatic Speech Recognition Although speech clips are extracted from movies based on their corresponding subtitle timestamps, discrepancies in duration and clarity may arise, especially in complex dialogue scenes and extended speeches. In addition, denoising speeches can sometimes distort human voices, making them challenging to recognize amidst background noise. To

[4] https://docs.fileformat.com/video/srt/
[5] https://www.nikse.dk/subtitleedit
[6] https://imsdb.com/
[7] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[8] https://github.com/resemble-ai/resemble-enhance

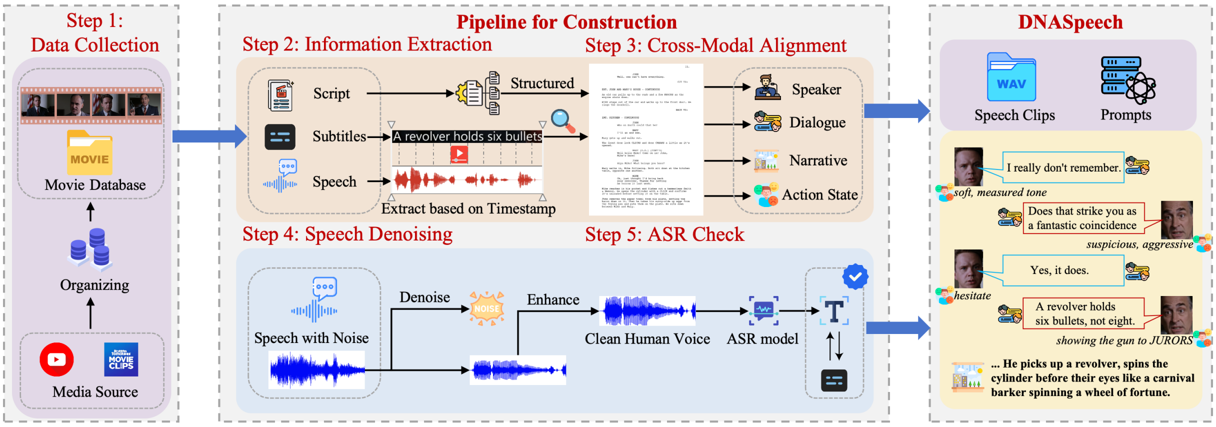


Figure 3: The automatic annotation pipeline for DNASpeech consists of five fundamental steps: (1) data collection of movie materials, (2) information extraction of textual content, (3) cross-modal alignment among "DNA" prompts, text, and speech, (4) speech denoising to reduce background noises and (5) automatic speech recognition to ensure the speech quality. An illustrative example from DNASpeech is provided on the right side.

ensure the quality and accuracy of the extracted speeches, it is necsssary to verify them against two criteria: (1) their recognizability and (2) alignment between their content text and the corresponding subtitles. We employ Automatic Speech Recognition (ASR) technology and make the reasonable assumption that if a speech clip can be accurately transcribed by an ASR model, it can also be recognized by humans. We use OpenAI's whisper-large-v3[9] for automatic speech recognition. Samples that do not match their corresponding subtitles after the ASR transcription are eliminated. With this validation process, we finish the construction pipeline of DNASpeech, ensuring its integrity and reliability for subsequent research.

3.3 Manual Assessment

After a series of rigorous filtering and screening processes in the pipeline, the quality of samples in DNASpeech generally meets our requirements. Next, further manual assessment is implemented to ensure the high quality of the data and consistency in the subjective evaluation of multiple evaluators. We manually evaluate each sample and assign scores ranging from 1 to 3 based on the overall quality of the sample. The specific criteria for scoring include (1) clarity; (2) emotional richness; (3) speech speed, avoiding excessively fast or slow pacing and (4) the relevance of the speech to the contextual information. Evaluators first score the samples based on each criterion independently, disregarding the other factors. Subsequently, we aggregate the evaluators' scores to obtain an overall quality assessment of each sample and the mean evaluation score for DNASpeech is 2.57. For detailed information about the evaluators, please refer to Appendix H.1.

3.4 Data Quality Verification

Although the primary purpose of DNASpeech is to aid in CS-TTS task, its inherent text-to-speech mappings make it also suitable for general TTS tasks. Therefore, we can verify its quality by examining the performance of DNASpeech on general TTS tasks. To demonstrate this, we select two TTS models: Tacotron2 and FastSpeech2, along with our baseline model DNA-TTS. Besides, we choose LJSpeech (Ito and Johnson, 2017) and DailyTalk (Lee et al., 2023) as the comparison datasets. For DNASpeech, we first clustered the data by speaker, then randomly sampled 90% of the examples from each speaker for the training set, with the remaining 10% forming the test set. By comparing the performance of these models on DNASpeech with their performance on the comparison datasets, we can assess the effectiveness of DNASpeech as a general TTS dataset.

Following the same setting as DailyTalk, we use mean opinion score (MOS) test as our evaluation metrics. MOS requires evaluators to rate the overall quality of the speech from 1 to 5, with higher scores representing better quality. Three listeners participated in the evaluation process, each holding a master's degree and having completed prior training. After each round of testing, we calculate the Kendall's W coefficient for the scores provided by

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

the three listeners. The results are accepted only when the Kendall's W coefficient ≥ 0.5, ensuring consistency in the ratings. Results in Table 1 show that models trained on DNASpeech sound as natural as those trained on other datasets, which proves the data quality of DNASpeech.

Model	LJSpeech	DailyTalk	DNASpeech
GT	4.07 ± 0.08	3.97 ± 0.07	4.05 ± 0.08
Tacotron2	3.87 ± 0.09	3.85 ± 0.10	3.90 ± 0.07
FastSpeech2	3.98 ± 0.07	3.97 ± 0.08	4.01 ± 0.07

Table 1: TTS integrity test result for DNASpeech. Score from 1 to 5. A higher score indicates better speech quality. GT refers to the speeches converted from ground truth mel-spectrograms.

4 Experiments

4.1 Existing Baselines

To evaluate the CS-TTS task, we select several representative text-to-speech methods as baselines for comparison. Based on the input data format and the architecture of models, we categorize these baselines into 3 types:

None-Prompt TTS, including Tacotron2 (Shen et al., 2018), FastSpeech2 (Ren et al., 2020), StyleTTS (Li et al., 2022b) and StyleSpeech (Min et al., 2021).

Prompt based TTS, including PromptTTS2 (Leng et al., 2023), PromptTTS++ (Shimizu et al., 2024), InstructTTS (Yang et al., 2024) and VoiceLDM (Lee et al., 2024).

Codec TTS, including VALL-E (Wang et al., 2023), NaturalSpeech2 (Shen et al., 2023) and VoiceCraft (Peng et al., 2024).

More details about these baselines are introduced in Appendix G.

4.2 Proposed Baseline

Since previous works are not tailored for the CS-TTS task, we design an intuitive baseline model to better evaluate the proposed benchmark. Our baseline model draws from the structure of PromptTTS (Li et al., 2022b) and consists of five main modules: Phoneme Encoder, Context Encoder, Style Fusion, Variance Adaptor, and Generator. Please refer to Appendix D for more details.

4.3 Leaderboard

To comprehensively evaluate baseline models' performance on CS-TTS benchmark, we use a combination of objective and subjective metrics.

4.3.1 Objective Metrics

Since ground truth waveform is available, following (Wang et al., 2023; Peng et al., 2024), we use four different objective metrics: MCD (Kubichek, 1993), F0, WER and PESQ (Rix et al., 2001). Please refer to Appendix E for detailed definitions.

4.3.2 Subjective Metrics

CS-TTS with Narratives Previous work has been limited by the form of prompts, typically only considering prompts that directly describe speech and lacking the ability to utilize environment information (Guo et al., 2023; Leng et al., 2023; Yang et al., 2024). Therefore, we propose CS-TTS with narratives as our first benchmark. We maintain the same training and testing sets as mentioned in Chapter 3.4. For each sample, its environment description is adopted as the input prompt.

To better assess speech quality, our MOS evaluations focus on different aspects: MOS-E emphasizes the alignment of the speech with the environment description, including volume, timbre, and conveyed emotion, aiming to test the ability to utilize information within the environment description. MOS-C focuses on the consistency of the speech itself, with the goal of evaluating the stability of the model when generating speech with the environment description. Please refer to Appendix H.2 for detailed evaluation guidelines.

CS-TTS with Dialogues Although previous work has explored the use of dialogue to control speech generation (Li et al., 2022a; Guo et al., 2021; Liu et al., 2023), they primarily focus on the content of the dialogue itself, neglecting the influence of the conversational scenario (e.g., the speaker's actions and expressions). Therefore, we propose CS-TTS with dialogues, which utilizes the speaker's action states as supplementary information to simulate the scenario of live conversations.

We first use MOS-D to assess the coherence between the speech and the dialogue context. During the evaluation, we primarily consider two factors: the overall emotional tone of the dialogue and the content of the most recent dialogue turn. To evaluate the impact of the action states on the speech, we employ MOS-S to determine whether the speech aligns with the action states. In this assessment, evaluators are initially provided with the dialogue context and action states to infer the speech's emotion, pitch, volume, etc., before listening to the

Model	Narrative		Dialogue		Objective Metrics			
	MOS-E ↑	MOS-C ↑	MOS-D ↑	MOS-S ↑	PESQ ↑	MCD ↓	F0 ↓	WER ↓
None-Prompt TTS Models								
Tacotron2	3.86 ± 0.05	3.92 ± 0.09	3.73 ± 0.06	3.65 ± 0.07	3.67	8.25	76.29	10.10
FastSpeech2	3.84 ± 0.08	$\mathbf{3.97 \pm 0.13}$	3.75 ± 0.09	3.69 ± 0.09	3.49	8.45	78.26	11.94
StyleTTS	$\mathbf{3.92 \pm 0.11}$	3.93 ± 0.07	$\mathbf{3.78 \pm 0.07}$	$\mathbf{3.72 \pm 0.06}$	3.22	8.34	**69.57**	9.76
StyleSpeech	3.89 ± 0.08	3.90 ± 0.09	3.77 ± 0.09	$\mathbf{3.72 \pm 0.11}$	**3.70**	**8.06**	71.04	**8.63**
Prompt-based TTS Models								
PromptTTS2	3.93 ± 0.07	3.92 ± 0.11	3.83 ± 0.11	3.80 ± 0.07	3.89	7.92	72.77	8.02
PromptTTS++	3.93 ± 0.09	3.99 ± 0.10	3.78 ± 0.08	3.70 ± 0.09	3.68	7.82	74.59	8.69
InstructTTS	3.94 ± 0.09	$\mathbf{4.12 \pm 0.08}$	3.83 ± 0.13	3.75 ± 0.08	3.89	7.50	72.65	7.56
VoiceLDM	3.94 ± 0.07	3.86 ± 0.06	3.83 ± 0.09	3.72 ± 0.08	3.75	7.57	76.83	6.74
DNA-TTS (Ours)	$\mathbf{3.96 \pm 0.09}$	4.01 ± 0.13	$\mathbf{3.85 \pm 0.06}$	$\mathbf{3.83 \pm 0.07}$	**4.10**	**7.35**	**71.45**	**6.36**
Codec TTS Models								
VALL-E	3.89 ± 0.06	3.95 ± 0.09	3.76 ± 0.05	3.74 ± 0.09	4.27	7.39	67.05	6.40
NaturalSpeech2	3.92 ± 0.04	4.03 ± 0.07	3.82 ± 0.05	3.79 ± 0.06	**4.38**	7.47	**66.20**	6.22
VoiceCraft	$\mathbf{3.94 \pm 0.08}$	$\mathbf{4.16 \pm 0.10}$	$\mathbf{3.88 \pm 0.06}$	$\mathbf{3.89 \pm 0.07}$	4.18	**7.16**	68.90	**6.03**

Table 2: Leaderboard results of DNASpeech. MOS-E and MOS-C are metrics of CS-TTS with narratives. MOS-D and MOS-S are metrics of CS-TTS with dialogues. The best results are highlighted in **bold**.

generated speech. They then evaluate the degree of alignment between the two and provide a final score. Please refer to Appendix H.2 for detailed evaluation guidelines.

4.4 Discussions

The evaluation results are presented in Table 2. Based on the results, we find that:

MOS-E and MOS-C metrics are generally correlated. This correlation suggests that models adept at capturing and integrating environmental descriptions—such as volume, timbre, and conveyed emotion—tend to maintain a high degree of consistency in their speech generation. This alignment underscores the importance of robust environmental context integration mechanisms in TTS systems to achieve both expressive and reliable speech synthesis.

Prompt-based methods perform better in terms of MOS-D, highlighting the efficacy of incorporating dialogue context in speech synthesis. This improvement is likely attributable to the models' ability to leverage contextual information from preceding dialogue turns, thereby producing more contextually appropriate and emotionally resonant speech. This advantage underscores the importance of dialogue-aware mechanisms in TTS systems, particularly for applications requiring dynamic and context-sensitive interactions. We further explore the influence of dialogue turns in Appendix F.

Codec TTS Models lead in both subjective and objective evaluations. The superior performance of Codec TTS models can be attributed to their advanced encoding mechanisms, which effectively capture and reproduce intricate speech nuances, including prosody, intonation, and emotional subtleties. These sophisticated encoding strategies enable Codec TTS systems to generate speech that not only aligns closely with environmental and contextual descriptions but also maintains high fidelity and naturalness, thereby setting a benchmark for future advancements in text-to-speech technology.

5 Conclusion

In this work, we introduce Contextualized and Situated Text-to-Speech (CS-TTS), aiming to generate speech that adapts to its surrounding context. To address the limitations of existing datasets, we collected a new dataset called DNASpeech to facilitate the development of CS-TTS. This dataset contains high-quality speech recordings annotated with "DNA" prompts that consist of Dialogues, Narratives, and Actions.

Furthermore, we establish a leaderboard to compare the performance of various TTS models on the CS-TTS task and propose a baseline method to serve as a reference for future research in this area. The results indicate that incorporating contextual and situated information can further enhance the performance of TTS models. We believe that DNASpeech can drive progress in TTS research, moving toward generating smooth and natural speech without manual intervention.

Limitations

There are two main key aspects we aim to address in our future work. Firstly, DNASpeech collects speech data from movie scenes rather than from real-world scenarios, which might affect the characteristics of the speech. We plan to diversify our dataset by incorporating speech data from more varied and real-world contexts to better reflect authentic speech patterns. Additionally, although we define more comprehensive contextualized and situated prompts than previous TTS datasets, it does not cover all possible prompt types. We intend to explore and integrate additional types of textual prompts to further enrich the dataset, enhancing its utility for a wider range of TTS applications.

References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Xavier Anguera, Nestor Perez, Andreu Urruela, and Nuria Oliver. 2011. Automatic synchronization of electronic and audio books via tts alignment and silence filtering. In *2011 ieee international conference on multimedia and expo*, pages 1–6. IEEE.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haohan Guo, Shaofei Zhang, Frank K Soong, Lei He, and Lei Xie. 2021. Conversational end-to-end tts for voice agents. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 403–409. IEEE.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. 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.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Jodi Kearns. 2014. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021. Expressive text-to-speech using style tag. *arXiv preprint arXiv:2104.00436*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, pages 125–128. IEEE.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. 2024. Voiceldm: Text-to-speech with environmental context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12566–12571. IEEE.

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, et al. 2023. Prompttts 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*.

Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su. 2022a. Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7917–7921. IEEE.

Xiang Li, Zhi-Qi Cheng, Jun-Yan He, Xiaojiang Peng, and Alexander G Hauptmann. 2024. Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis. *arXiv preprint arXiv:2404.18398*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2022b. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *arXiv preprint arXiv:2205.15439*.

Yuchen Liu, Haoyu Zhang, Shichao Liu, Xiang Yin, Zejun Ma, and Qin Jin. 2023. Emotionally situated text-to-speech synthesis in user-agent conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5966–5974.

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

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, pages 7748–7759. PMLR.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. 2024. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 2024. 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)*, pages 12672–12676. IEEE.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

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*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE.

A License

The dataset [10] is available for free download and non-commercial use under the CC BY-NC-SA 4.0 license.

B Social Impact

Given the sensitive nature of biometric data, particularly vocal recordings, all data undergo anonymization to protect personal privacy. However, despite these measures, there exists a potential risk of misuse. To prevent unauthorized usage or dissemination, access to the dataset is subject to a rigorous review process. Regarding the intended use, users are permitted to define their own tasks in our dataset under the license, upon advanced contact with us.

C Statistics

We analyze the statistics of speeches, focusing on both pitch and speed to overall present DNASpeech. We extract the F0 fundamental frequency from speeches to obtain their pitch. As shown in Fig 4, the pitch distribution range for female speakers is wider than that for male speakers, evenly distributed from 70Hz to 150Hz; in contrast, the pitch for male speakers is more concentrated, mostly appearing in the 65Hz-95Hz range. Overall, the pitch of female speakers is generally higher than that of male speakers. To more accurately measure the speed of a speech, we calculate the syllables per second (SPS) after removing its silent segments. The distribution shown in the figure indicates that the speakers' speech speed ranges from 6 SPS to 22 SPS, with the 12-15 SPS being the most frequent.

D Proposed Baseline

D.1 Model Architecture

We propose a specific baseline for CT-TTS task, as shown in Fig 5. The Phoneme Encoder uses BERT (Devlin et al., 2019) to encode the phonemes of the speech. The Context Encoder shares the same structure as the Phoneme Encoder but includes classification tasks for emotion, pitch, energy, and speed during training. To ensure that the generated speech accurately reflects the contextualized and situated descriptions provided in the prompts, we introduce a Style Fusion module that employs a cross-attention mechanism for fine-grained feature fusion.

Given that prompts in the CS-TTS task do not include descriptions of acoustic features, we insert a speaker embedding into the fused representation to control the characteristics of the speech. Inspired by the setup of FastSpeech2 (Ren et al., 2020), we incorporate a Variance Adaptor module following the Style Fusion. This module predicts information such as duration, pitch, and loudness, further clarifying the speech characteristics and addressing the one-to-many problem in prompt-based TTS tasks. The final output of our baseline model is a mel-spectrogram, which is transformed into speech using a pre-trained HiFiGAN (Kong et al., 2020), ensuring high-fidelity speech synthesis.

D.2 Effect of Modules

In our proposed baseline (DNA-TTS), the Context Encoder and Style Fusion module collectively serve as the core dialogue-aware components. Specifically:

- **Classification Task of Context Encoder:** This module employs BERT to encode contextual features. More importantly, during training, it performs auxiliary classification tasks for emotion, pitch, and energy, enabling it to capture nuanced conversational cues (e.g., shifts in tone or intent across dialogue turns).
- **Style Fusion:** Leveraging cross-attention, this module dynamically aligns the encoded dialogue context with the current input phonemes. This ensures that synthesized speech reflects the inferred emotional trajectory and speaker intentions from prior dialogue turns, thereby improving coherence (MOS-D).

To quantify the impact of these two components, we add ablation experiments, where we progressively remove these two components during both training and inference stages. The results are as follows:

Stage	MOS-D	PESQ	MCD	F0	WER
Original Model	3.85	4.10	7.35	71.45	6.36
- CLS Task	3.80	3.86	7.78	72.37	7.78
- Style Fusion	3.74	3.59	8.29	74.38	8.03

Based on the experimental results, it can be observed that the model's performance gradually declines as components are disabled. Specifically, when only the classification task is removed, there

[10] https://anonymous.4open.science/r/DNASpeech-FDCD

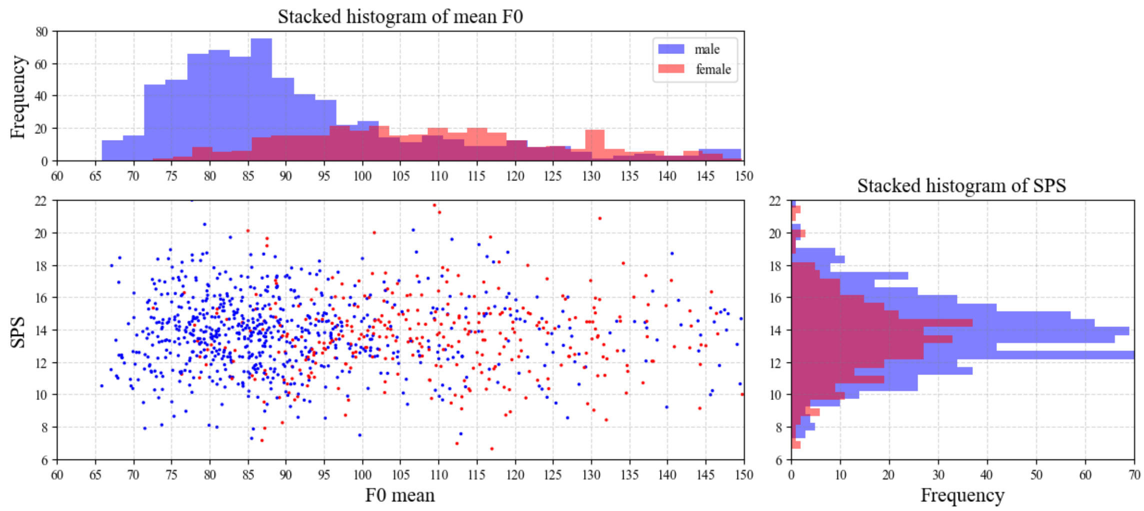
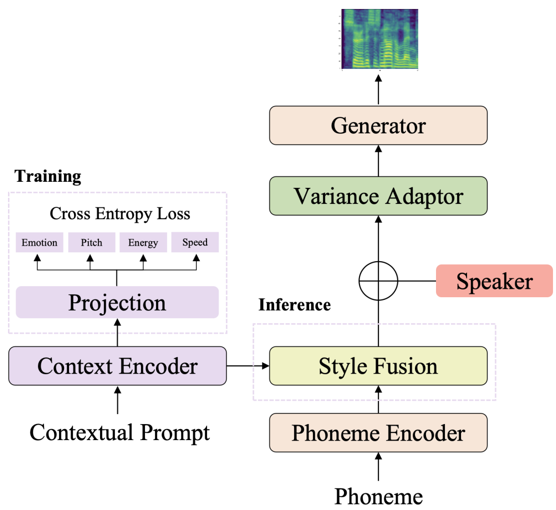


Figure 4: The statistical distribution of the mean F0 and SPS. Each point in the scatter figure represents a speaker. The top and right figures are stacked histograms of mean F0 and SPS by gender.



(a) DNA-TTS

Figure 5: Illustration of the architecture of the proposed baseline for CS-TTS tasks.

is a noticeable drop in performance. This may be because the contextual information was not supervised and aligned during training, leading to insufficient handling of detailed features such as emotion, pitch, and speed. When style fusion is further removed, the model's performance degrades to a level comparable to that of None-Prompt TTS models, at which point the contextual information can not be integrated with the text input.

E Definition of Objective Metrics

MCD (Mel-Cepstral Distortion) (Kubichek, 1993) measures the difference of Mel Frequency Cepstrum Coefficients (MFCC) between generated and ground truth, defined as

$$\mathrm{MCD} = \frac{10}{\ln 10}\sqrt{\frac{1}{2}\sum_{i=1}^{L}(m_i^g - m_i^r)^2}$$

where L is the order of MFCC, which we set to be 13. m_i^g is the i^{th} MFCC of ground truth recording and m_i^r is the i^{th} MFCC of the generated speech. We use the pymcd package [11] for calculating MCD.

F0 is measured by estimating the fundamental frequency of the audio and calculating the F0 distance between the grounding truth and the generated speech. A smaller F0 distance indicates that the generated speech is closer to the grounding truth. For F0 estimation, we use the pYIN algorithm implemented in librosa, with a minimum frequency of 65 Hz and a maximum frequency of 200 Hz.

WER (Word Error Rate) is used to measure the difference between the predicted and actual transcription of speech by calculating the minimum number of substitutions, deletions, and insertions required to change the system's output into the reference text:

$$\mathrm{WER} = \frac{S + D + I}{N}$$

where S refers to substitutions, D refers to deletions, I refers to insertions and N is the total number of words in the reference transcription. We use whisper-large-v3 [12] as our ASR model.

[11] https://github.com/chenqi008/pymcd
[12] https://huggingface.co/openai/whisper-large-v3

PESQ (Perceptual Evaluation of Speech Quality) (Rix et al., 2001) is an objective metric developed by the International Telecommunication Union (ITU) in recommendation P.862 and is commonly used for evaluating the quality of speech in telecommunication systems, such as voice over IP (VoIP) and TTS. It models the human auditory system's perception of speech. We use the pesq package [13] for calculating PESQ.

F Influence of Dialogue Turns

To assess the impact of contextual information quantity on speech quality, we conduct additional experiments. Specifically, we further divided the DNASpeech test set into four categories based on the number of dialogue turns: **1-3 turns**, **4-6 turns**, **7-8 turns**, and **8 turns or more**. We then test both DNA-TTS (Prompt-based TTS Models) and VALL-E (Codec TTS Models) on these subsets, and the results are shown in Table 3:

The results show that contextual information has a positive effect on speech quality within a certain range, with the model performance typically peaking around the 4-6 dialogue turns. However, as the number of dialogue turns increases, the speech quality begins to decline. When the contextual information becomes too lengthy (i.e., beyond 8 turns), the speech quality significantly deteriorates. This may be due to the contextual information becoming too dispersed, losing its supervisory effect on speech generation. This serves as a reminder to be cautious when using contextual information to avoid such issues.

G Baseline details

G.1 Introduction of Baselines

Tacotron2 (Shen et al., 2018) leverages an end-to-end deep learning framework, where the input is a sequence of text and the output is a spectrogram, which is then used to generate natural-sounding speech. The model uses a sequence-to-sequence architecture with attention mechanisms, allowing it to learn a direct mapping between textual features and audio characteristics.

FastSpeech2 (Ren et al., 2020) designed to enhance the efficiency, reliability, and flexibility of speech synthesis systems. Unlike traditional autoregressive models that generate audio sequentially, FastSpeech employs a non-autoregressive architecture, enabling parallel generation of speech outputs. Additionally, FastSpeech incorporates mechanisms to improve robustness against input variations and allows for greater controllability over speech characteristics such as prosody and intonation.

PromptTTS2 (Leng et al., 2023) incorporates a variation network that predicts voice variability not captured by text prompts, and a prompt generation pipeline that leverages large language models (LLMs) to compose high-quality text prompts automatically. The variation network in PromptTTS 2 works by predicting the representation from reference speech based on the text prompt representation, allowing for the sampling of diverse voice variability.

PromptTTS++ (Shimizu et al., 2024) designed to synthesize the acoustic characteristics of various speakers based on natural language descriptions. This method employs an additional speaker prompt to efficiently map natural language descriptions to the acoustic features of different speakers.

PromptTTS++ (Shimizu et al., 2024) builds upon the concept of prompt-based TTS, where voice characteristics can be manipulated through descriptive prompts. A key innovation in PromptTTS++ is the introduction of "speaker prompts", which are designed to describe voice attributes like gender-neutral, young, old, and muffled, and are intended to be independent of speaking style. To facilitate this, the authors constructed a dataset based on the LibriTTS-R corpus with manually annotated speaker prompts, as no large-scale dataset with such annotations existed. The system employs a diffusion-based acoustic model along with mixture density networks to capture diverse speaker characteristics from the training data.

InstructTTS (Yang et al., 2024) is designed to synthesize speech with varying speaking styles by using natural language as style prompts. This model introduces an insightful approach to controlling the expressiveness of synthetic speech, such as emotion and speaking rate, through natural language descriptions, which can include detailed instructions. It models acoustic features in a discrete latent space, using a discrete diffusion probabilistic model to generate vector-quantized (VQ) acoustic tokens instead of the traditional mel spectrogram.

StyleSpeech (Min et al., 2021) is designed to generate high-quality, personalized speech for multiple speakers with minimal audio samples from the target speaker. This model is particularly adept at adapting to new speakers with short-duration audio samples. StyleSpeech introduces a novel Style-

[13] https://github.com/ludlows/PESQ

Model	Turns	MOS-D↑	MOS-S↑	PESQ↑	MCD↓	F0↓	WER↓
DNA-TTS	1-3	3.87	**3.85**	4.16	**7.03**	69.53	**6.03**
	4-6	**3.89**	**3.85**	**4.23**	7.25	**68.90**	6.29
	7-8	3.84	3.83	4.12	7.52	71.45	6.43
	>8	3.80	3.79	3.89	7.60	75.87	6.69
VALL-E	1-3	3.77	**3.78**	4.28	7.45	**66.69**	**6.28**
	4-6	**3.79**	3.76	**4.31**	**7.38**	67.19	6.45
	7-8	3.73	3.73	4.24	7.62	67.75	6.55
	>8	3.68	3.65	4.15	7.94	68.27	6.72

Table 3: The performance of DNA-TTS and VALL-E using different dialogue turns. The best results are highlighted in **bold**, while the worst results are marked with underline.

Adaptive Layer Normalization (SALN) technique that aligns the text input's gain and bias according to the style extracted from a reference speech audio. This allows the model to synthesize speech in the style of the target speaker effectively.

StyleTTS (Li et al., 2022b) focuses on generating natural and diverse speech. StyleTTS is designed to overcome the challenges of producing speech with realistic prosodic variations, speaking styles, and emotional tones. A key innovation of StyleTTS is the integration of style-based generative modeling into a parallel TTS framework, which allows it to synthesize speech that captures the stylistic nuances of reference audio. This is achieved through the use of a novel Transferable Monotonic Aligner (TMA) and duration-invariant data augmentation, enhancing the model's ability to produce speech with natural prosody and speaker similarity.

VoiceLDM (Lee et al., 2024) sets a new standard in audio generation by incorporating environmental context into the synthesis process. Unlike traditional TTS models that focus solely on linguistic content, VoiceLDM is designed to respond to two types of natural language prompts: a description prompt that outlines the environmental setting of the audio, and a content prompt that specifies the linguistic content of the speech.

VALL-E (Wang et al., 2023) represents a significant shift in the approach to TTS. Unlike traditional methods that treat TTS as a continuous signal regression problem, VALL-E frames TTS as a conditional language modeling task. This model leverages discrete codes derived from an off-the-shelf neural audio codec model, which allows it to synthesize high-quality, personalized speech with minimal acoustic prompts. VALL-E outperforms existing state-of-the-art zero-shot TTS systems in terms of speech naturalness and speaker similarity. Additionally, VALL-E is capable of preserving the speaker's emotion and acoustic environment in the synthesized speech.

NaturalSpeech2 (Shen et al., 2023) aims to synthesize natural and human-like speech with high quality and diversity. NaturalSpeech 2 employs a neural audio codec that converts speech waveforms into sequences of latent vectors and a diffusion model that generates these vectors based on text input. A key feature of NaturalSpeech 2 is its zero-shot capability, which allows the system to synthesize diverse speech even for unseen speakers, demonstrating superior prosody/timbre similarity, robustness, and voice quality compared to previous TTS systems.

VoiceCraft (Peng et al., 2024) is a token-infilling neural codec language model that excels in both speech editing and zero-shot text-to-speech applications. VoiceCraft is designed to work with various audio sources, including audiobooks, internet videos, and podcasts. It utilizes a Transformer decoder architecture and employs a unique token rearrangement process that combines causal masking and delayed stacking. This innovative approach allows the model to generate speech that is nearly indistinguishable from original recordings in terms of naturalness, as evaluated by human listeners.

G.2 Training Parameters

Training parameters are listed in Table 4 and Table 5.

H Evaluation Details

H.1 Evaluator Information

A total of eight evaluators participated in the manual evaluation process of this work. All evaluators

Model	Optimizer	β_1	β_2	ϵ	Batch size	Training steps	Learning rate
Tacotron2	Adam	0.9	0.99	10^{-6}	16	2 epochs	10^{-4}
FastSpeech2	Adam	0.9	0.98	10^{-9}	16	2 epochs	10^{-5}
StyleTTS	AdamW	0	0.99	10^{-7}	16	2 epochs	10^{-4}
StyleSpeech	Adam	0.9	0.98	10^{-9}	16	2 epochs	2×10^{-4}
PromptTTS2	Adam	0.9	0.99	10^{-7}	16	2 epochs	10^{-5}
PromptTTS++	Adam	0.9	0.99	10^{-7}	16	2 epochs	10^{-5}
InstructTTS	AdamW	0.9	0.94	10^{-7}	16	2 epochs	3×10^{-6}
VoiceLDM	AdamW	0.9	0.99	10^{-7}	16	2 epochs	2×10^{-5}

Table 4: Training configurations for different models

Model	Schedule	Other params
Tacotron2	/	/
FastSpeech2	Linear schedule	Warm up step=200
StyleTTS	OneCycleLR	Weight decay=10^{-4}, $\lambda_{s2s} = 0.2$, $\lambda_{adv} = 1$, $\lambda_{mono} = 5$, $\lambda_{fm} = 0.2$, $\lambda_{dur} = 1$, $\lambda_{f0} = 0.1$, $\lambda_n = 1$
StyleSpeech	/	/
PromptTTS2	/	/
PromptTTS++	/	/
InstructTTS	Linear schedule	Warm up step=200
VoiceLDM	/	Drop rate of c_{desc}=0.1, Drop rate of c_{cont}=0.1

Table 5: Training configurations for different models

held a graduate degree or higher, including three individuals of Asian descent and five native English speakers. Prior to the evaluation, all participants were thoroughly briefed on the evaluation methods and specific guidelines.

H.2 Guidelines

H.2.1 MOS-E

Purpose. MOS-E evaluates how well the system's speech aligns with the environment description, taking into account volume, timbre, and the emotion conveyed. The focus is on how effectively the system incorporates the environmental context into its speech, ensuring that the output feels contextually appropriate, emotionally consistent, and well-matched to the described surroundings.

Criteria.

1. Volume Appropriateness: Does the system adjust its volume in a way that matches the described environment? For instance, if the environment is a quiet room, is the speech soft and subtle? If the setting is a loud, bustling street, does the system compensate with louder or more intense speech?

2. Timbre Alignment: Does the system adjust the tone or texture of its voice to fit the environment? For example, in a serene setting like a forest, is the voice calm and soothing, whereas in a high-energy environment like a sports stadium, does the voice reflect excitement or intensity?

3. Emotion Conveyed: Is the emotional tone of the speech consistent with the environment description? If the environment is described as tense or somber (e.g., a dark alley or a funeral), does the speech reflect that tension or sadness? If the environment is happy or lively, does the voice convey a matching positive emotion?

4. Contextual Adaptation: How well does the system integrate information from the environment description into its speech output? Does the system fully utilize the given context, or does it fail to adapt its voice appropriately?

Scoring Instructions.

1. Very Poor (1): The speech is completely out of sync with the environment description. Volume, timbre, and emotion are inappropriate, making the system's output feel disconnected from the described surroundings.

2. Poor (2): The speech shows some effort to match the environment but is still significantly mismatched. There may be a lack of emotional depth or incorrect volume/timbre adjustments that detract from the immersion.

3. Moderate (3): The speech aligns to some degree with the environment, but it is inconsistent. Volume and timbre might be correct in some cases, but emotional expression or contextual adaptation could be improved.

4. Good (4): The speech is generally well-aligned with the environment. Volume, timbre, and emotion are appropriately adjusted most of the time, with only minor discrepancies.

5. Excellent (5): The system's speech perfectly matches the environment description. It seamlessly adjusts volume, timbre, and emotion to create a highly immersive and contextually accurate experience.

Considerations. Evaluate the system's ability to adapt dynamically to the environmental cues. Pay attention to the subtlety of the system's voice adjustments: a high-quality system should be able to make these adjustments in a natural, unobtrusive way that enhances the realism of the interaction.

H.2.2 MOS-C

Purpose. MOS-C evaluates the consistency of the system's speech when generating responses based on the given environment description. The focus is on assessing how stable the system is in maintaining a steady and coherent output throughout the interaction, ensuring that the speech remains consistent in terms of tone, style, and quality, regardless of environmental shifts or changes in the context.

Criteria.

1. Tone Consistency: Is the system's tone consistent throughout the interaction? Does the system maintain a coherent style (e.g., formal, informal, casual, etc.) without unnecessary fluctuations in tone?

2. Volume Stability: Does the system keep a stable volume level during the interaction? Even if the environment description changes, is there an appropriate, but consistent, volume level maintained without abrupt changes?

3. Timbre Consistency: Is the timbre (quality of the voice) stable and consistent across multiple turns? Does it retain its distinct characteristics, or does it fluctuate in a way that feels unnatural?

4. Emotional Consistency: Does the system maintain a stable emotional tone, or does it randomly fluctuate? Emotional shifts should occur only when the environment changes in a way that justifies them (e.g., a shift from a happy environment to a sad one).

5. Stylistic Continuity: Does the system maintain consistency in its speaking style, such as formality, in line with the environment description?

Scoring Instructions.

1. Very Inconsistent (1): The system's speech is highly unstable, with frequent and noticeable shifts in tone, volume, timbre, or emotion that do not match the environment or create a jarring user experience.

2. Inconsistent (2): There are noticeable fluctuations in the speech output that disrupt the flow of the interaction. These shifts may seem unnatural or out of place in the context of the environment.

3. Moderately Consistent (3): The system maintains an overall stable speech output, but some inconsistencies are present. There may be occasional fluctuations in tone or volume, but they don't significantly impact the coherence of the speech.

4. Consistent (4): The speech remains fairly stable throughout the interaction, with minor inconsistencies that do not detract from the overall experience. The system maintains an appropriate tone, volume, timbre, and emotional consistency.

5. Highly Consistent (5): The system's speech is completely stable, with no noticeable fluctuations. Tone, volume, timbre, and emotion remain coherent and aligned with the environment description throughout the entire interaction.

Considerations. Evaluate the uniformity of the speech characteristics. A consistent system will adapt to environmental changes subtly without sudden shifts that could break immersion or distract from the user experience.

H.2.3 MOS-D

Purpose. MOS-D evaluates the coherence of the system's speech in relation to the ongoing dialogue context. The goal is to assess how well the system's responses align with the previous conversation history and whether they maintain logical flow and relevance. This score focuses on the system's ability to stay on-topic, build on prior exchanges, and provide responses that are contextually appropriate within the dialogue.

Criteria.

1. Relevance to Previous Turns: Does the system's response directly address the most recent user input? Are there clear connections to prior exchanges, or does the response seem disconnected or out of place?

2. Logical Flow: Does the system's speech follow a natural progression from previous dialogue? Are responses structured in a way that makes sense given what has been discussed so far?

3. Turn-taking and Timing: Does the system understand and respect the natural flow of conversation, responding at appropriate moments and allowing for smooth turn-taking? Does it avoid interrupting or providing responses that feel out of sync with the timing of the conversation?

Scoring Instructions.

1. Very Incoherent (1): The system's response is completely disconnected from the previous dialogue. It may ignore or misunderstand the context, resulting in responses that feel irrelevant or random.

2. Incoherent (2): The response is partially relevant but lacks clear connection to the ongoing conversation. There are significant gaps in logical flow or misunderstandings of the context.

3. Moderately Coherent (3): The system's response is somewhat relevant, but there may be minor lapses in coherence. It addresses the user's input, but the response could be more fluid or better integrated with the context.

4. Coherent (4): The response is mostly relevant and logically follows from previous turns. There are minor inconsistencies, but the overall flow of the conversation is maintained.

5. Highly Coherent (5): The system's response is seamlessly integrated into the ongoing dialogue. It builds naturally on previous exchanges, remains relevant, and maintains a logical and smooth conversational flow throughout.

Considerations. Pay close attention to how well the system recognizes the dialogue history and context. The response should not only be appropriate to the immediate previous turn but also reflect understanding of the overall direction of the conversation. A highly coherent system will effectively navigate and build on the evolving dialogue while keeping responses consistent and contextually relevant.

H.2.4 MOS-S

Purpose. MOS-S evaluates how well the system's speech aligns with the action states described in the given dialogue context. This assessment focuses on determining whether the speech accurately reflects the inferred emotion, pitch, volume, and other relevant qualities based on the action states provided to the evaluator. The goal is to assess the system's ability to generate speech that is consistent with the intended emotional tone, energy level, and contextual cues.

Criteria.

1. Emotion Alignment: Does the speech accurately reflect the emotion inferred from the action states and dialogue context? For example, if the action state indicates anger, is the speech delivered with an appropriate intensity and emotional weight?

2. Pitch Consistency: Is the pitch of the speech consistent with the emotional tone and action state? A heightened pitch might be expected for excitement, while a lower pitch may suit a calm or serious environment.

3. Volume Appropriateness: Does the volume of the speech align with the inferred action state? For example, if the action state suggests an intense or confrontational situation, should the speech be louder or more forceful, as opposed to a quiet, subdued volume for a calm or intimate setting?

4. Timbre Alignment: Is the timbre (quality of the voice) consistent with the action states? For example, a high-energy situation might require a brighter, more vibrant voice, while a somber situation could demand a more muted, heavy tone.

Scoring Instructions.

1. Very Poor (1): The speech is completely mismatched with the action states. Emotion, pitch, volume, and timbre are completely off, making the generated speech feel disconnected from the described action states.

2. Poor (2): The speech has some attempt at matching the action states, but significant discrepancies exist. The emotional tone, volume, or pitch are not fully aligned with the intended action states, resulting in a noticeable mismatch.

3. Moderate (3): The speech aligns moderately well with the action states. There are some noticeable differences in emotion, pitch, volume, or timbre, but the overall speech still corresponds with the intended context and action states.

4. Good (4): The speech is generally well-aligned with the action states. The emotion, pitch, volume, and timbre are mostly consistent with the inferred context, with only minor inconsistencies.

5. Excellent (5): The speech perfectly aligns with the action states. The emotion, pitch, volume, and timbre are precisely matched to the action states and the overall dialogue context, enhancing the realism and immersion of the interaction.

Considerations. Evaluate how well the system translates the inferred action states (emotion, volume, pitch, etc.) into speech characteristics. The more closely the generated speech matches the action states, the higher the alignment score. Pay special attention to subtle aspects like the emotional tone and how the system handles shifts in the action state across different turns.