Computational Narrative Understanding for Expressive Text-to-Speech

Gaspard Michel[†*]
gmichel@deezer.com

Elena V. Epure[†◇]
elena.epure@idiap.ch

Christophe Cerisara[*]
christophe.cerisara@loria.fr

[†] Deezer Research, Paris, France
[*] Loria, Nancy, France
[◇] IDIAP, Martigny, Switzerland

Abstract

Recent advances in text-to-speech (TTS) have been driven by large, multi-domain speech corpora, yet the expressive potential of audiobook data remains underexamined. We argue that human-narrated audiobooks, particularly fictional works, contain rich and diverse prosodic cues arising from the natural alternation between neutral narration and expressive character dialogue. Building from this observation, we introduce LibriQuote, a large-scale 5.3K hours of expressive speech drawn from character quotations. Each quote is supplemented with contextual pseudo-labels for speech verbs and adverbs that characterize the intended delivery of direct speech (e.g., "*he whispered softly*"). We found that fine-tuning a flow-matching model on LibriQuote yields substantial improvements in expressivity and intelligibility, while training from scratch enhances expressiveness of an autoregressive TTS model. Benchmarking on LibriQuote-*test* highlights significant variability across systems in generating expressive speech. We publicly release the dataset, code, and evaluation resources to facilitate reproducibility. Audio samples can be found at https://libriquote.github.io/.

1 Introduction

Recently, text-to-speech (TTS) systems have significantly improved by scaling to larger, multi-domain speech corpora (Shen et al., 2023; Wang et al., 2023; Kharitonov et al., 2023; Jiang et al., 2023a; Anastassiou et al., 2024; Wang et al., 2024; Chen et al., 2024; Wang et al., 2025), exhibiting remarkable naturalness, quality and voice following abilities. Typically, these large-scale corpora contain speech derived from *audiobooks* (Kahn et al., 2019; Pratap et al., 2020; Kang et al., 2024) or in-the-wild data (He et al., 2024). While the latter may include diverse expressivity and speaking styles, it is sometimes claimed that audiobooks may lack such expressiveness (He et al., 2024).

Contrary to this, we argue that human readings of audiobooks, especially fictional works, offer extensive prosodic variability through a natural alternation between neutral speech and expressive utterances. Neutral speech usually corresponds to narrative parts, the fragments enunciated by the narrator, while expressive speech can be employed during dialogues between fictional characters. Dialogue is an essential narrative function (Genette, 1983) that allows fictional characters to portray themselves and express aspects of their personality, emotional states, and world representation. As the story unfolds, characters navigate dynamically through their own emotion arc (Vishnubhotla et al., 2024), which can be portrayed during audiobook reading. Indeed, numerous blog posts written by professional narrators emphasize the importance of *impersonation*, or the act of bringing life to fictional characters (Gonzalez, 2021; Browm, 2021; Goodwin, 2024).

In this work, we put fictional characters at the forefront, by creating *a large-scale dataset of expressive quotes* and *proving its effectiveness for expressive TTS*. Our dataset, **LibriQuote**, differs from standard large-scale audiobook datasets (Pratap et al., 2020; Kahn et al., 2019; Kang et al., 2024) by relying on a narrative-aware segmentation focused on character quotes that exhibit expressivity and diversity, as opposed to narration, which corresponds to a more neutral reading style. We automatically annotate, and validate with humans, speech cues specific to quotations as contextual information: pseudo-labels of *speech verbs* and *adverbs* used in the narrative to characterize direct speech utterances (*e.g.* "he whispered softly"). LibriQuote-*train* includes 3300 speakers and 5.3K hours of quotations and 12.7K hours of narration, while LibriQuote-*test* provides 7.4 hours of quotations from 15 speakers.

Our findings are the following:

Findings of the Association for Computational Linguistics: ACL 2026, pages 6194–6215

1. *Fine-tuning* a state-of-the-art autoregressive TTS system with LibriQuote-*train* does not directly improves expressivity, but yields *substantial improvements in speech intelligibility* on in-domain and out-of-domain data (Section 4). Besides, when *training from scratch* the same TTS system with LibriQuote-*train*, we *increase speech expressivity*, but at the cost of reduced intelligibility due to substantially reduced training data (Ye et al., 2025). In contrast, *fine-tuning a flow-matching model directly enhances both speech expressivity and intelligibility*. Finally, we show that *using both narrations and quotations* contained in LibriQuote-*train* is *a promising direction for training expressive and intelligible TTS systems*.

2. The benchmarking of various TTS systems on LibriQuote-*test* reveals large discrepancies in TTS systems' ability to synthesize expressive speech (Section 5). This indicates that LibriQuote-*test* represents a challenging benchmark that could foster advances in developing expressive TTS.

The dataset is available at the following links:
Huggingface: https://huggingface.co/datasets/gasmichel/LibriQuote
GitHub: https://github.com/deezer/libriquote

2 Background and Motivation

2.1 Related Speech Datasets

Audiobooks have long been the de-facto open-source data for training TTS systems. The LibriVox project records thousands of public-domain books read by volunteer speakers. LibriSpeech (Panayotov et al., 2015), a 1000-hour English corpus covering 2400 speakers and extracted from LibriVox, has been widely used to train and benchmark TTS systems. LibriTTS (Zen et al., 2019) offers an enhanced version of LibriSpeech with better audio quality and utterance segmentation, though with less training data, while LibriSpeech-PC (Meister et al., 2023) restores punctuation and capitalization in LibriSpeech transcripts.

Libri-Light (Kahn et al., 2019) is the largest open English audiobook corpus at about 60,000 hours of LibriVox speech but lacks audio transcriptions. The MLS dataset (Pratap et al., 2020) includes 50,000 hours of multilingual audiobooks and provides transcriptions aligned with the original text. LibriHeavy (Kang et al., 2024) improves Libri-Light by aligning audio segments with their corresponding book text, but does not provide contextual book information around segments. These 30-second segments are cut at sentence boundaries and may include multiple quotations and narration, alternating between neutral (narration) and expressive (quotation) speech.

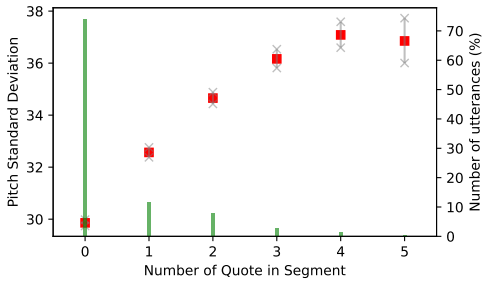

Figure 1: Average pitch standard deviations (red squares) per number of quotations in audio segments in the LibriHeavy-*small*, with bootstrapped 95% confidence intervals (grey lines); and percentage of total number of segments per number of quotations (green).

Moreover, several efforts towards the creation of expressive read speech corpora have been made. Emotional speech datasets (Busso et al., 2008; Cao et al., 2014; Livingstone and Russo, 2018; Zhou et al., 2021) that assign discrete emotion labels to speech utterances have been extensively used to evaluate emotional synthesize of TTS systems (Jiang et al., 2023b; Wang et al., 2024, 2025). The L2-ARCTIC corpus (Zhao et al., 2018) proposes an 11 hour corpus with ten speakers, spanning 5 different accents. EXPRESSO (Nguyen et al., 2023) introduces an high-quality dataset of read and improvised speech across 26 expressive styles.

2.2 Gaps in Current Datasets

Prior works that use audiobooks as a resource for speech data, such as LibriTTS or LibriHeavy, have overlooked aspects of narrative discourse. They either completely discard character quotations, or combine them with other quotations and/or neutral narrative parts in their audio segments, which are constructed from random sentence breaks. Such segments are thus likely to exhibit multiple "distinct" pitch distributions, possibly alternating between neutral, emotional and/or pitch-varying speech. We illustrate this effect in Figure 1.

Specifically, we performed quotation detection

on 120000 segments from LibriHeavy-*small*, and calculated their pitch standard deviation in the corresponding audio segments. We found that 75% of segments are only narrative parts, while the remaining 25% contain from 1 to 12 quotations. We see that pitch standard deviation typically increases with the number of quotations within the segment (Spearman $\rho = 0.218$, $p < 0.001$).

Modeling these complex pitch distributions is a challenging task. It can hinder the training of TTS systems, which may end up focusing on learning to model the simpler, narrative parts instead. By shifting the focus from formal narration to character quotations, LibriQuote provides varied direct speech samples that span a broader range of emotions, at scale. These traits are desirable for generating high-quality, human-like speech (He et al., 2024). Moreover, we supplement these quotations with narrative information indicating *how* an utterance should be spoken, providing a natural language pseudo-label for expressivity.

3 The LibriQuote Dataset

Distinguishing between "story" (what is happening) and "discourse" (how it is told) is a key aspect of narrative understanding. NLP research has extensively focused on various aspects of *narrative discourse*, as theorized by Genette (1983). Recently, a variety of works explored the usage of Large Language Models (LLMs) to model elements of narrative understanding, such as dialogue (Michel et al., 2025), character roles and information extraction (Stammbach et al., 2022; Gurung and Lapata, 2024), narrativity (Hatzel and Biemann, 2024) or structure (Soni et al., 2023).

The current work is grounded in the perspectival issues of fictional characters, including point of view and dialogues, anchored in the *voice* linking function of Genette's theoretical framework. In fiction, utterances are sometimes supplemented with *narrative information* in the form of speech verbs and adverbs (*e.g. "he whispered softly"*). These indicators have been shown to produce positive effects during reading, such as enhancing reading comprehension (Wolters et al., 2022) or character identification (Van Krieken et al., 2017). For audiobook narrators, they are a valuable resource to set the tone and point of view of the character. Other contextual information such as the emotional context of a character are likely to also impact speech delivery, but are often prone to subjective interpretation and are harder to automatically infer (Vishnubhotla et al., 2024).

3.1 Data Creation Pipeline

Our data creation pipeline is similar to the one proposed for LibriHeavy (Kang et al., 2024). However, we filter out books that are unlikely to contain fictional characters, such as bibliographical or other types of non-fictional works. We detail the filtering and audio-text alignment stages below.

3.1.1 Audio Preparation

We start by browsing the LibriVox API[1] to collect recording information. We only extract English recordings corresponding to Fiction books. This step is necessary to filter out non-fiction books that are unlikely to contain fictional characters. Each recording is downsampled to 16kHz for further processing, but we provide links to download the original audio files as well, allowing LibriQuote to support systems operating on higher audio quality. LibriVox also includes *Dramatic Readings*, where each fictional character is performed by a different speaker. We exclude these recordings from LibriQuote, as they would require an additional speaker diarization step.[2]

3.1.2 Text Preparation

For each audio recording, we download the corresponding book text from Project Gutenberg. Some books have external text providers, but manual inspection revealed that automatically downloading from these websites was often too complicated or yielded low-quality texts, frequently due to OCR errors. In such cases, we turned to Project Gutenberg to search for the target book and discarded it if it was not available on the platform. We then perform quotation detection using BookNLP[3] on each book. We discard every book containing less than 20 quotations, as we found that these books typically yield quotation detection errors. For other books, we keep quotation positions in the original text for further alignments.

3.1.3 Audio Transcription

We start by segmenting each audio recording into 30-second segments, with a 2 second overlap at each side. These segments are then transcribed with

[1] https://librivox.org/api/info

[2] We applied the same alignment process to these recordings and plan to release these alignments in future work.

[3] https://github.com/booknlp/booknlp/blob/main/booknlp/english/litbank_quote.py

	Train			Dev	Test
	N	**Q**	$\mathbf{Q}_f$	**Q**	**Q**
Count	3,87M	3,51M	378K	2921	5598
Hours	12723	5359	379	5.1	7.4
Dur (s)	11.8	5.5	3.6	6.2	4.8
Speakers	3314		2818	56	15
avg (h)	3.8	1.6	0.13	0.1	0.5
Books	2991		2900	65	27

Table 1: Descriptive statistics of LibriQuote. Count is the utterance number. **N** refers to narration, **Q** to quotations and $\mathbf{Q}_f$ to filtered quotation ($\mathbf{Q}_f \subset \mathbf{Q}$).

a Zipformer-Transducer ASR model trained on LibriSpeech[4], and combined by leveraging word-level timestamps to obtain a full timestamped transcript of the recording.

3.1.4 Text-Audio Alignment

Since audio recordings often involve a full chapter, this step is necessary to align the corresponding chapter text within the full original book. This alignment process involves two stages. In the first stage, *close matches* between a transcription and the book text are constructed. These *close matches* correspond to (i, j) pairs where i is an index in the transcription and j an index in the book text. Then, a coarse alignment is produced by taking the longest chain of pairs $(i_1, j_1), \ldots, (i_N, j_N)$, such that $i_1 \leq \cdots \leq i_N$ and $j_1 \leq \cdots \leq j_N$. The second stage produces a final alignment by concatenating Levenshtein alignments (Lcvenshtcin, 1966) between the transcription and the text segment produced by the longest chain of pairs. We refer the reader to Kang et al. (2024) for additional information regarding the alignment process.

3.1.5 Quotation Alignment

The output of the text-audio alignment is an array of words in the original text, timestamped in the audio recording. We leverage the timestamps and the output of the quotation detection system to segment each audio into a set of read quotations. Additionally, we build audio segments from narration paragraphs (*i.e.* parts that are not quotations) in each book, and trimmed long start end silences for each audio segment. The final construction stage involved assigning contextual information around quotations. Based on the alignments, we match each quotation with a window of around 100 words occurring before and after it, with boundaries defined by paragraphs. With LibriQuote, we release the preprocessed texts as well as full paragraph and quotation alignments. Thus, it is possible to derive a contextual window of any desirable length around quotations.

[4]https://github.com/k2-fsa/icefall/pull/1058

3.2 Dataset Characteristics

Table 1 shows LibriQuote's descriptive statistics.

Training Set The filtering and alignment process resulted in a total of 12723 training hours of narration and 5359 training hours of quotation across 2991 unique books and 3314 unique speakers. As expected, narration parts constitute the majority of audiobook recording in terms of duration, but quotations still represent almost 50% of the total number of training utterances. Indeed, the average audio segment duration is 11.8 seconds for narration parts and 5.5 seconds for quotations. LibriVox does not provide gender information of its readers, so the balance between male and female speaker in the training set is unclear.

Dev & Test Sets We designed the dev and test sets to evaluate whether synthesized speech in quotations have the correct level of expressiveness. Each quotation serves as the expressive speech, and we match them with the nearest narration utterance from the book context that acts as the neutral reference speech. Synthesizing these quotations can be particularly challenging, as the text alone may be misleading with regard to the true emotion, which is often specified by the surrounding context. As an example, the exclamation mark in the quotation *"It's so good!" he whispered softly* might suggest it shall be uttered with a loud voice, but the narrative information clarifies that it is actually being whispered. Thus, we also provide contextual information in the vicinity of quotations to encourage further research on predicting expressivity from textual cues for audiobook TTS. The test set contains 8 male and 7 female unseen speakers, for a total of 7.5 hours.

The dev set contains random utterances from speakers and books in the train set. It thus does not encompass zero-shot speech synthesis, but aims at evaluating expressive synthesis from neutral speech for seen speakers. However, we ensure no speaker-overlap with the *test* set.

Test Set Analysis We start the exploration of LibriQuote-*test* by computing Word Error Rate (WER) whith `Whisper-Large v3` and an auto-

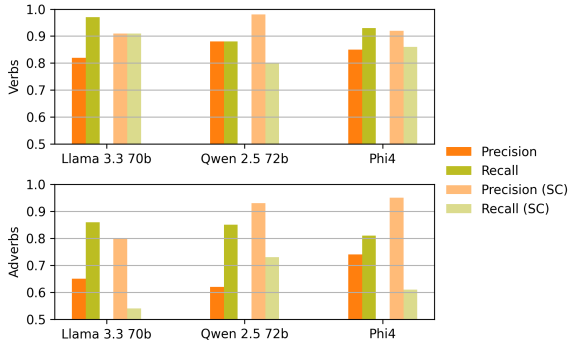

Figure 2: Evaluation of the extraction of Verbs (top) and Adverbs (bottom). SC indicates self-confidence.

mated metric for speech naturalness, UTMOS (Saeki et al., 2022), ranging from 1 (unnatural) to 5 (very natural). We found that utterances exhibiting high WER often contain only a few words and proper names such as "*Hey, Jean*". We thus decided to remove 205 utterances from the test set that contain less than two words. Other segments with high WER were sometimes spoken with a strong accent, which might affect Whisper transcription, but we decided to keep these segments. UTMOS analysis reveal an average value of 3.48, where most lower valued utterances occur when a quotation is spoken with very high emotion intensity. We thus chose not to filer out these utterances, as they correspond to desired expressivity.

To further analyse LibriQuote test set, we explore which emotions are conveyed in quotations and reference narrations. Following Wang et al. (2024), we compute emotion representations using Emotion2Vec[5] (Ma et al., 2024). We found that LibriQuote-*test* quotations convey a larger set of predicted emotions compared to their narration reference and LibriHeavy segments: only 67% of quotations are predicted as neutral against 87% and 91% for the reference and LibriHeavy respectively. Further details can be found in Appendix A.

3.3 Contextual Cues for Direct Speech

We extract narrative cues indicating the speaking style of a quotation (described by the narrator), using a contextual window spanning one paragraph before and after the quotation. For this, we rely on several LLMs as they have shown promising results in literary dialogue understanding (Piper and Bagga, 2024; Michel et al., 2025).

We use few-shot prompting with 5 examples. To mitigate extraction errors and keep only informative narrative content, we replace all quotations in context with special markers, such that only narration and structure remains. Our experiments involved prompting to self-report a *confidence score* on a 1-to-10 scale that we used to prune-out unconfident predictions in order to maximize precision. More details are provided in Appendix B.

Validation In an initial phase dedicated to developing guidelines and solving complex cases, two annotators independently tagged 100 quotations. After discussion and once satisfactory agreement was reached (we report a Cohen's κ score of 0.87 for verbs and adverbs), a single annotator continued with 300 extra instances, yielding a total of 400 quotations annotated with *verbs*, *adverbs*, *nouns*, and *adjectives*, drawn from the train split. We found that adjectives and nouns were relatively rare in the annotated data, and that LLMs failed entirely to extract them. We thus report results only for verbs and adverbs in Figure 2.

We see that most models perform relatively well at extracting narrative cues, but still suffers from hallucinations, in particular for adverbs. When pruning out all predictions with a confidence score lower than 10, all models exhibit higher precision at the cost of lower recall for both verbs and adverbs. A high precision is desirable: we want to avoid quotations to be tagged with wrong expressive speech verbs and adverbs. We thus select Phi-4 as the base LLM for further extraction of contextual cues on the full corpora, as it obtained the highest precision score on adverbs and a satisfactory precision on verbs, while also being the smallest and fastest model.

High-Expressivity Split After extracting contextual cues on the full dataset with the strategy described above, we build a high-expressivity subset. First, we include all quotations that had a non-empty adverb pseudo-label regardless of the predicted speech verb, as adverbs often give precise information on how a quotation is uttered (*e.g.* "*he said cautiously*"). Then, based on a manually defined set of *expressive* speech verbs[6], we add all quotations that have a verb falling into that list. The resulting split, $\mathbf{Q}_f$, contains 377,776 quotations (11% of the full quotation train set) for a total of 379 hours, which may be used for data-efficient expressive TTS, or can foster automatic analysis of

[5] https://huggingface.co/emotion2vec/emotion2vec_plus_base

[6] The full list of verbs is available in Appendix B.2.

	WER	SIM-O	E-Sim	CtxMOS	WR
GT	6.5	-	-	3.55	-
SparkTTS	4.8	0.46	0.69	2.94	38 %
FT ($\mathbf{Q}_f$)	**4.6**	0.47*	0.71	2.97	41%
FT ($\mathbf{Q}$)	5.9*	0.46*	0.71	2.89	37%
Scratch ($\mathbf{Q}$)	9.5*	0.40*	0.71	3.09**	38%
+ Ctxt	8.2*	0.40*	0.71	3.15*	35%
Full ($\mathbf{N} \cup \mathbf{Q}$)	5.1*	0.41*	0.71	3.30*	37%
+ FT ($\mathbf{Q}_f$)	5.1*	0.41*	0.71	3.30*	**41%**
F5-TTS	6.9	0.53	0.71	2.95	31 %
FT ($\mathbf{Q}_f$)	6.6*	**0.54***	0.71	**3.33***	26 %

Table 2: Evaluation on LibriQuote-*test* set of various training setups. Two-sided paired student t-test are conducted against SparkTTS or F5-TTS (* indicates $p < 0.05$, and ** indicates $p < 0.1$).

	LibriSpeech PC		SeedTTS *test-en*	
	WER ↓	SIM ↑	WER ↓	SIM ↑
Ground Truth	2.44	0.69	2.06	0.73
SparkTTS	3.06	**0.52**	2.64	0.46
FT ($\mathbf{Q}_f$)	2.10*	0.51	2.07	0.42*
FT ($\mathbf{Q}$)	**2.00***	0.51	1.90*	0.42*
Scratch ($\mathbf{Q}$)	3.27	0.47*	5.14*	0.41*
Full ($\mathbf{N} \cup \mathbf{Q}$)	2.22*	0.49*	4.05*	0.41*
+ FT ($\mathbf{Q}_f$)	2.49*	0.49*	4.23*	0.42*
F5-TTS	2.11	0.66	1.69	0.67
FT ($\mathbf{Q}_f$)	2.13	0.66	**1.64**	**0.67**

Table 3: Word Error Rate (WER) and speaker similarity (SIM) on LibriSpeech-PC and Seed-TTS *test-en*. Two-sided paired student t-tests are conducted against SparkTTS (* indicates $p < 0.05$).

the alignment between written narrative cues and uttered speech.

4 Training Experiments

In this section, we analyse the impact of using LibriQuote as a resource for fine-tuning or training TTS systems from scratch. Our goal is to understand 1) how LibriQuote could be used as an alternative to standard TTS datasets such as Emilia (He et al., 2024), even though it contains 20× less data and 2) how it can be used to enhance expressivity of pre-trained TTS systems when used as a fine-tuning resource. We mainly conduct experiments with SparkTTS and F5-TTS, that we chose as baselines for better reproducibility. SparkTTS is an autoregressive TTS model based on Qwen2-0.5B, that generates global and semantic discrete audio tokens that can be converted back to raw audio. It was trained on approximately 100K hours of speech data, containing in-the-wild and audiobook corpora. F5-TTS is a non-autoregressive flow-matching based TTS system, trained on 100K hours of speech data. We describe SparkTTS and F5-TTS in details in Appendix F. For each models, we use the available checkpoints and do not modify generation parameters. For SparkTTS, we compute global and semantic tokens for each training utterances, and fine-tune the LLM-backbone using the standard language modeling task on semantic tokens with text and global tokens prepended in the sequence. For F5-TTS, we use the provided fine-tuning script available in the official repository[7]. Training details including optimizer and hyperparameters can be found in Appendix C and Appendix D.

We design training experiments in two ways: fine-tuning experiments with varying quotation dataset (either the full set of quotations $\mathbf{Q}$ or the expressive subset $\mathbf{Q_f}$), and training from scratch experiments with either the full quotation subset $\mathbf{Q}$ or the full training dataset (both narrations and quotations $\mathbf{N} \cup \mathbf{Q}$). Besides, we experiment with replacing SparkTTS text-condition with contextual information in the vicinity of a target quotation (denoted as `Ctxt`). By having access to more contextual information, we expect the second variant to improve expressivity over standard SparkTTS. More details can be found in Appendix C.

4.1 Metrics

Our evaluation follows the *cross-sentence* design (Le et al., 2023): a narration utterance of 2 to 15 seconds is used as audio context to guide the generation of a quotation utterance.

Objective Metrics We report WER computed with Whisper-large-v3 (Radford et al., 2023) to measure speech intelligibility and speaker similarity between the synthesized speech and the original ground-truth speech (SIM-O). We employ a WavLM-large based speaker verification model (Chen et al., 2022) to extract speaker embeddings and calculate cosine similarities. We evaluate emotion similarity (E-Sim) by leveraging representations from Emotion2Vec (Ma et al., 2024)

[7] https://github.com/swivid/f5-tts

(`emotion2vec-plus-base`). Similarity scores are calculated by computing the cosine similarity between the ground-truth and synthesized speech representations.

Contextual Metrics To measure to which extent synthesized speech matches the quotation context, we leverage recent advances in Large-Audio-Language-Models (LALMs) and use an LALM-as a-Judge approach (Manku et al., 2025; Ji et al., 2025). We follow Manku et al. (2025) and use Gemini-2.5 Pro as the Judge model, as it was proved to correlate strongly with human judgements. We build an expressive test-subset containing 201 quotations from LibriQuote-*test*, sampling only quotations that are matched with a predicted adverb, following the methodology described in Section 3.3. We evaluate synthesized and ground-truth utterances in two ways: a single sample Context Mean Opinion Score (ContextMOS) and a two-sample comparative Win-Rate score as defined in Manku et al. (2025). ContextMOS rates on 1-5 scale how an utterance speaking style matches the book context, while Win-Rate calculates the percentage of times a synthesized utterance delivers a more appropriate speaking style than a ground-truth sample, based on the book context. Details can be found in Appendix E.

4.2 Benchmark Datasets

We leverage standard benchmark datasets to evaluate in-domain and out-of-domain TTS performance. We use LibriSpeech-PC (Meister et al., 2023), a punctuation restored version of LibriSpeech with the standard split from Chen et al. (2024) that contains 1127 samples with 4-to-10 seconds audio prompts and SeedTTS *test-en* (Anastassiou et al., 2024), which is composed of 1088 samples from Common Voice (Ardila et al., 2020). We report WER and cosine similarity between synthesized speech and reference sample (SIM) with the same models as described above.

4.3 Analysis

Table 2 displays results on LibriQuote-*test*. The ground-truth WER is relatively high compared to standard audiobook datasets such as LibriSpeech. As discussed in Section 3.2, this is partly due to strong emotions or accents in some samples, which increases recognition errors. In contrast, SparkTTS and its variants fine-tuned with $\mathbf{Q}_f$ and $\mathbf{Q}$ achieve lower WER than the ground truth. These fine-tuned models show slightly higher SIM-O and E-Sim, while ContextMOS and win rates remain similar, suggesting that fine-tuning on LibriQuote modestly improves intelligibility and speaker similarity but not contextual expressiveness. Conversely, models trained from scratch show the opposite trend: using standard text conditions yields a ContextMOS of 3.09 when using the quotation subset as training data, which increases to 3.15 when book context is used as input. This improved expressivity comes with reduced intelligibility and speaker similarity, likely due to a limited amount of training data. This decrease in intelligibility is drastically reduced when using the both narrations and quotations ($\mathbf{N} \cup \mathbf{Q}$) when training from scratch. This model achieves greater expressivity compared to all SparkTTS baselines with a ContextMOS of 3.30. Interestingly, fine-tuning this model on a second-stage with the expressive subset $\mathbf{Q_f}$ does not yield significant gains, echoing the other fine-tuning experiments.

Compared to SparkTTS, we found that fine-tuning conducted with F5-TTS is able to yield large expressivity gain. When fine-tuning with the filtered quotation subset $\mathbf{Q_f}$, F5-TTS achieves better speech intelligibility and improved expressivity, with ContextMOS reaching 0.4 points higher. This notable difference is likely related to the different training objectives: while SparkTTS uses an autoregressive loss, it seems that the flow-matching loss of F5-TTS is better able to pick the important expressive information present in LibriQuote's quotations.

Benchmark Datasets In Table 3, we present results for benchmark datasets. We omit the from-scratch model with context because these datasets lack additional contextual information. SparkTTS results are reported using samples generated by our synthesis pipeline to ensure fair comparison. Unlike LibriQuote-*test*, fine-tuned models achieve lower WER than SparkTTS on LibriSpeech-PC and SeedTTS-*test-en*, indicating better intelligibility. We attribute this to fine-tuning acting as a data curriculum, where high-quality data is introduced late in training, a strategy used in recent TTS training strategies (Atamanenko et al., 2025). On LibriSpeech-PC, SIM is comparable between SparkTTS and FT models, but degrades on the out-of-domain SeedTTS-*test-en*. The Scratch model shows similar WER to SparkTTS on LibriSpeech-PC but performs much worse on SeedTTS, as ex-

	WER ↓	SIM-O ↑	E-Sim ↑	ContextMOS	Win-Rate	CMOS	MOS
Ground Truth	6.5	-	0.62*	3.55 ±0.20	-	0.0	3.58 ±0.15
SparkTTS	**4.8**	0.46	0.69	2.94 ±0.20	38%	-0.97 ±0.25	3.26 ±0.15
F5-TTS	6.9	0.53	0.71	2.95 ±0.21	31%	-0.98 ±0.25	3.47 ±0.16
MaskGCT	7.6	**0.56**	**0.72**	2.94 ±0.20	28%	-0.98 ±0.22	3.36 ±0.16
IndexTTS2	5.2	0.49	0.63	3.33 ±0.19	46%	**0.25** ±0.42	3.47 ±0.29
IndexTTS2-*Context*	5.4	0.50	0.64	**3.45** ±0.21	**54%**	0.03 ±0.45	**3.63** ±0.29

Table 4: Zero-shot TTS results on LibriQuote-*test* set. Similarity metrics are computed against Ground Truth (* indicates similarity against reference). Bootstrapped 95% confidence intervals are reported with ±.

pected given the mismatch between audiobook and in-the-wild speech. The SparkTTS model trained on the full dataset ($\mathbf{N} \cup \mathbf{Q}$) largely improves on speech intelligibility and similarity over the other from-scratch variant, but still struggles on out-of-domain data. Further fine-tuning this model on the expressive subset $\mathbf{Q_f}$ slightly decreases speech intelligibility on both datasets, further suggesting that fine-tuning is not necessary for a model already trained on the full LibriQuote dataset.

These experiments conclude that supervised fine-tuning on LibriQuote significantly enhances SparkTTS intelligibility on in-domain and out-of-domain data, and that training from scratch enhances expressivity at the cost of less intelligible speech. Besides, fine-tuning a different TTS backbone yields a different picture, showing improves expressivity and intelligibility for F5-TTS. This difference highlights important research directions in understanding the reasons behind less performant SparkTTS fine-tuned models.

Thus, LibriQuote appears as a promising resource to fine-tune flow-matching based TTS systems to be more expressive, or to train from-scratch autoregressive TTS models for an enhanced expressivity.

5 Benchmarking with LibriQuote-*test*

In this section, we evaluate the ability of state-of-the-art TTS systems to synthesise speech as expressive as amateur audiobook readers from LibriQuote-*test*. We focus our evaluation on aspects of speech naturalness and emerging capacities to systematically deliver speech utterances that satisfy an implicit or explicit speech intent (e.g. "whispering" or "happiness"). We compare four recent open-source TTS systems: SparkTTS (Wang et al., 2025), F5-TTS (Chen et al., 2024), MaskGCT (Wang et al., 2024) and IndexTTS2 (Zhou et al., 2025). These models were chosen based on their availability and performance on standard benchmarks, and are described in Appendix F. Notably, IndexTTS2 first predicts a 7-class emotion distribution based on an emotion prompt (with the text to synthesize as default), that conditions the synthesized speech. This decoupled approach is well-suited for LibriQuote, as the emotion prediction can be done on contextual information around quotations rather than quotation-text alone. We denote this approach as IndexTTS2-*Context*. All models were trained with the Emilia dataset (He et al., 2024) on approximately 50K to 100K hours of speech data, but SparkTTS and IndexTTS2 also use additional Audiobook speech data. We use publicly available checkpoints for each model and do not modify generation parameters.

5.1 Metrics

We use the same set of objective and contextual metrics defined in Section 4.1. In addition, we conduct a subjective rating with human raters. We employ Mean Opinion Score (MOS, rated from 1 to 5 with 0.5 intervals) to assess naturalness, and Comparative MOS (CMOS, rated from -3 to 3) to measure the degree of expressivity with respect to the ground-truth. We randomly selected 2 samples per speaker in LibriQuote-*test*, for a total of 30 samples. Each sample was judged by 5 raters. More details can be found in Appendix G.

5.2 Analysis

Results are displayed in Table 4. SparkTTS has the lowest WER (4.8) but the lowest SIM-O, while MaskGCT shows the highest WER and SIM-O. F5-TTS and MackGCT produce speech with relatively high emotion similarity. Both IndexTTS2 variants achieve low WER and high ContextMOS and Win-Rates, with IndexTTS2-*Context* matching ground truth in Win-Rate (54%). Subjective tests further show IndexTTS2 reaches positive CMOS, meaning it is rated more expressive than ground truth on average. Other TTS systems sound natural (high

MOS) but fail to capture appropriate prosody, as indicated by negative CMOS and low ContextMOS, highlighting IndexTTS2's clear advantage in expressive speech synthesis.

While IndexTTS2 already captures emotional expressivity well, replacing text conditioning with contextual conditioning for emotion prediction further improves contextual metrics, achieving human-level Win-Rate ($\geq$50%) and MOS naturalness. These results show that decoupling emotion prediction from speech synthesis significantly enhances contextual expressivity. Although IndexTTS2 generates more expressive speech than other baselines, it also yields the lowest E-Sim, likely due to emotion prediction errors from text alone. Incorrect emotion predictions can produce speech that contradicts the intended style, which might mislead listeners in applications such as Audiobook reading. As detailed in Appendix H, Quotations read by amateur speakers do not always convey sufficient emotion, but still render the appropriate style more often than all models. In contrast, IndexTTS2 exhibits a higher proportion of both appropriate and mismatched speaking styles than other models, an issue partially mitigated by IndexTTS2-Context.

6 Conclusion

We proposed LibriQuote, a public speech dataset derived from audiobooks, that includes 12,720 hours of neutral narration and 5,359 hours of expressive quotations from fictional characters read by amateur readers. Our qualitative analysis reveals that LibriQuote-*test* is more emotionally diverse than previous audiobook corpora. Besides, each utterance of LibriQuote is supplemented with its book context and pseudo-labels of quotation intent, showing a wide-range of intended speaking style. Fine-tuning experiments show both training subsets improve SparkTTS intelligibility on other corpora while training from scratch on the full quotation set enhances expressiveness, and that fine-tuning F5-TTS on LibriQuote's expressive subset yields substantial expressivity gains. Evaluation on LibriQuote-*test* shows that IndexTTS2 matches ground-truth expressivity, though some synthesized quotations convey unintended emotions. In-depth analysis revealed that amateur readers sometimes convey insufficient emotion in their rendering of quotations, but TTS systems do so at an even higher proportion. Despite recent advancements, TTS systems require substantial improvements to meet standards of professional audiobook narration.

We also want to highlight LibriQuote's value for practitioners of Digital Humanities. LibriQuote provides 3 million character quotations in text and speech modalities, from 2,991 public domain novels, with pseudo-labels of quotation intent. This large-scale data offers new perspectives to analyse Audiobook prosody, e.g. comparing how male/female speakers interpret male/female characters (Pethe et al., 2025) or understanding stylistic differences of words uttered by characters (Michel et al., 2024).

7 Limitations

This work bears several limitations. First, we did not filter LibriQuote-*train* utterances that exhibit very high WER between the transcript and ground-truth text as done in previous work. Thus, LibriQuote might contain a small number of outliers, which can hinder learning of ASR or TTS systems. We plan to perform WER-based filtering in a future dataset release.

Second, our evaluation of expressivity relies on an LALM-as-a-judge approach, which is inherently biased by the LALM understanding abilities and own cultural biases. While prior works have incorporated such approaches in their evaluation pipelines, we believe these automated results should mostly be used to compare models against each other rather than make individual conclusions about models. Our findings, however, reveal closely aligned trends between human and model judgments, indicating that IndexTTS2 is a substantially more capable TTS system for producing expressive speech.

In addition, although our experiments focused directly on end-to-end TTS, we still note that LibriQuote ; and in particular its filtered subset ; might be well-suited for the training of Neural Audio Codecs. Indeed, numerous Audio Codecs targeting English language use LibriSpeech or LibriTTS as training data (Wang et al., 2025; Zhang et al., 2024). We thus believe that exploring the use of LibriQuote as training data for Neural Audio Codecs offers an interesting area of research for higher-fidelity reconstruction of expressive speech.

Finally, we experimented with narrative context conditioning in utterance synthesis, but we believe further research should be conducted on how TTS systems may benefit from it.

8 Ethical Considerations

This work is purely dedicated to research projects. While TTS models continue bridging the gap towards human naturalness and expressiveness, numerous dangerous applications arise, such as voice spoofing, or the forced replacement of talented voice actors by AI narrators. Therefore, we believe it is of paramount importance to ensure responsible and ethical applications of TTS.

We also note that LibriQuote does not provide explicit gender or accent distributions of its speakers. As a results, TTS models trained with LibriQuote are likely to exhibit gender or accent bias ; where one gender or one/multiple accents are either generated more faithfully or generated with stereotypical cues. We think researchers and/or companies using LibriQuote as training data in the purpose of creating a commercial tool should systematically assess and, if possible, mitigate such bias before providing users with an end product.

The human study involved participants recruited on a voluntarily basis from our organization, and we ensure each listening test was completed within working hours.

9 Acknowledgements

This work was performed using HPC resources from GENCI–IDRIS (Grant AD011011668R5)

References

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

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Oleg Atamanenko, Anna Chalova, Joseph Coombes, Nikki Cope, Phillip Dang, Zhifeng Deng, Jimmy Du, Michael Ermolenko, Feifan Fan, Yufei Feng, Cheryl Fichter, Pavel Filimonov, Louis Fischer, Kylan Gibbs, Valeria Gusarova, Pavel Karpik, Andreas Assad Kottner, Ian Lee, Oliver Louie, and 13 others. 2025. Tts-1 technical report. *Preprint*, arXiv:2507.21138.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

James Browm. 2021. The Art of Storytelling – How Voice Over Brings Audiobooks to Life.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE.

G. Genette. 1983. *Narrative Discourse: An Essay in Method*. Cornell paperbacks. Cornell University Press.

Luis Daniel Gonzalez. 2021. How to become an audiobook narrator.

Brittany Goodwin. 2024. So You Want To Narrate Audiobooks.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Alexander Gurung and Mirella Lapata. 2024. CHIRON: Rich character representations in long-form narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8523–8547, Miami, Florida, USA. Association for Computational Linguistics.

Hans Ole Hatzel and Chris Biemann. 2024. Story embeddings — narrative-focused representations of fictional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, Siqi Zheng, Jin Xu, Junyang Lin, and Zhou Zhao. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *Preprint*, arXiv:2505.09558.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, and 1 others. 2023a. Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis. *arXiv preprint arXiv:2307.07218*.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, and 1 others. 2023b. Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis. *arXiv preprint arXiv:2307.07218*.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. 2019. Libri-light: A benchmark for ASR with limited or no supervision. *CoRR*, abs/1912.07875.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. *Preprint*, arXiv:2309.08105.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023a. High-fidelity audio compression with improved rvqgan. *Preprint*, arXiv:2306.06546.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023b. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

VI Lcvenshtcin. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, volume 10.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and 1 others. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.

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

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

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. 2025. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *Preprint*, arXiv:2505.23009.

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models. *Preprint*, arXiv:2310.02943.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *Preprint*, arXiv:2309.15505.

Gaspard Michel, Elena Epure, Romain Hennequin, and Christophe Cerisara. 2024. Distinguishing fictional voices: a study of authorship verification models for quotation attribution. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 160–171, St. Julians, Malta. Association for Computational Linguistics.

Gaspard Michel, Elena V. Epure, Romain Hennequin, and Christophe Cerisara. 2025. Evaluating LLMs for quotation attribution in literary texts: A case study of LLaMa3. In *Proceedings of the 2025 Conference*

of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 742–755, Albuquerque, New Mexico. Association for Computational Linguistics.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, and 1 others. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

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.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. *Preprint*, arXiv:2211.03929.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Preprint*, arXiv:2212.09748.

Charuta Pethe, Bach Pham, Felix D Childress, Yunting Yin, and Steven Skiena. 2025. Prosody analysis of audiobooks. In *2025 19th International Conference on Semantic Computing (ICSC)*, page 217–221. IEEE.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 37–46, Miami, Florida, USA. Association for Computational Linguistics.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

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

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

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*.

Hubert Siuzdak. 2024. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *Preprint*, arXiv:2306.00814.

Sandeep Soni, Amanpreet Sihra, Elizabeth Evans, Matthew Wilkens, and David Bamman. 2023. Grounding characters and places in narrative text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11723–11736, Toronto, Canada. Association for Computational Linguistics.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Kobie Van Krieken, Hans Hoeken, and José Sanders. 2017. Evoking and measuring identification with narrative characters–a linguistic cues framework. *Frontiers in psychology*, 8:1190.

Krishnapriya Vishnubhotla, Adam Hammond, Graeme Hirst, and Saif Mohammad. 2024. The emotion dynamics of literary novels. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2557–2574, Bangkok, Thailand. Association for Computational Linguistics.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Alissa P Wolters, Young-Suk Grace Kim, and John William Szura. 2022. Is reading prosody related to reading comprehension? a meta-analysis. *Scientific Studies of Reading*, 26(1):1–20.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *Preprint*, arXiv:2502.04128.

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. *CoRR*, abs/1904.02882.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech large language models. *Preprint*, arXiv:2308.16692.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Interspeech 2018*, pages 2783–2787.

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.

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

A LibriQuote Analysis

As mentioned in Section 3.2, we performed emotion detection with the open source Emotion2Vec-plus-base. We report the detection results in Table 5 and display t-SNE projections of the resulting utterance representations in Figure 3, and compare LibriQuote-*test* quotations, narrations and LibriHeavy segments. As expected, we see a much higher proportion of quotations predicted as non-neutral in LibriQuote quotations compared to LibriHeavy and narration segments.

	Quotations	Narrations	LibriHeavy
Neutral	66.9%	86.8%	91.4%
Angry	4.6%	0.5%	0.8%
Sad	13.6%	9.3%	4.6%
Happy	8.4%	2.7%	2.3%
Surprised	4.5%	0.3%	0.6%
Fearful	0.9%	0%	0.1%
Disgusted	1.1%	0.4%	0.3%

Table 5: Proportion of emotion labels predicted by Emotion2Vec. We excluded *Unknown* and *Other* categories.

B Narrative Information

B.1 Prompting Experiments

We display in Figure 6 the prompt used to extract narrative information for each quotation. Note that we masked in-context quotations with special markers, and replace the target quotation with "[TARGET]". This is to ensure that only narrative elements remain in the contextual surroundings of a target quotation.

Additionally, we tested various Large Language Models (LLMs) for the task of narrative information extraction, and also tested the extraction of other elements such as *nouns* and *adjectives*. Full results can be found in Table 6. Nouns and adjectives often occur when the utterance description is supplemented by additional information (*e.g.* "he said with a *sigh*" for nouns and "he added in a *loud*, *sharp* tone" for adjectives). Note that to ensure a high-quality filtering, we tried to maximize the precision of predictions in order to avoid potential hallucinations where some narrative information are spuriously associated to utterances. We found that Qwen2.5-72b with self-confidence (SC) provides the best overall performance for these two types of information, followed closely by Phi-4 with SC. In the end, we decided to use Phi-4 with SC to extract information on the full train set, as it obtained very high precision in 10× less computation time than Qwen2.5-72b with SC. Besides, we found that LLMs dramatically fail at extracting accurately adjectives and nouns, with very low F1 scores. We found during the annotation that extracting such information is harder, even for humans, as introducing adjectives and nouns is less grounded in linguistic rules, and authors might use a more diverse set of linguistic devices to introduce these information. Thus, we decided not to rely on the extraction of such elements.

Overall, extracting verbs and adverbs with Phi-4 with SC (we used vLLM (Kwon et al., 2023) as the inference backend) on the full train set took around 6 hours on a single NVIDIA H100 GPU card.

B.2 Filtering

The result of the extraction is a (potentially empty) set containing extracted verbs and adverbs from Phi-4_{conf}. To produce the filtered subset $\mathbf{Q}_f$, we leverage independently the extracted verbs and adverbs in the following order:

1. We include all quotations that have a non-empty adverb.
2. Using the remaining quotations, we include all utterance that have an extracted verb that fall into a predefined list of speech verbs, $\mathbf{S}$.
3. We discard every other quotations.

To build the list of speech verbs, $\mathbf{S}$, we started by extracting a large list of 201 potential speech verbs from the web[8] along with their descriptions. Then,

[8] https://archiewahwah.wordpress.com/speech-verbs-list/

(a) LibriQuote (Quotation) (b) LibriQuote (Narration) (c) LibriHeavy (Mixed)

Figure 3: t-SNE projections of utterance representations computed with `emotion2vec-plus-base`.

based on verb descriptions, we discarded every verb that might indicate a *neutral* way of speaking. The resulting verb list contains 89 speech verbs and can be found in Table 7.

C SparkTTS Experiments

Our experiments described in Section 4 involved fine-tuning and training from scratch numerous variants of SparkTTS with LibriQuote. We provide details below for complete reproducibility of our experiments. Whether for fine-tuning or training from scratch, we transform each quotation utterance in the sequence $[T, G, S]$ where $T = (t_1, \ldots, t_N)$ are the text tokens, $G = (g_1, \ldots, g_{32})$ are SparkTTS global tokens (with a fixed number of 32 tokens) and $S = (s_1, \ldots, S_D)$ are semantic tokens, with varying length depending on the utterance length. Each training experiments involve optimizing the following language modeling task:

$$L = \frac{1}{D} \sum_{i=1}^{D} p(s_i | t_1, \ldots, t_N, g_1, \ldots, g_{32}, s_{j<i})$$

in other words, we optimize the Language Model backbone solely on semantic tokens.

C.1 Fine-Tuning

For both fine-tuning experiments, we employ Adam optimizer (Kingma and Ba, 2017) with peak learning rate of 1e-5. We train each model for 3 epochs on their respective data split ($\mathbf{Q}$ and $\mathbf{Q}_f$). A cosine schedule is used with 10000 warmup steps for the model fine-tuned on $\mathbf{Q}$ (FT($\mathbf{Q}$)) and 5000 steps when using $\mathbf{Q}_f$ (FT($\mathbf{Q}_f$)). Note that $\mathbf{Q}_f$ contains $10\times$ less utterances than $\mathbf{Q}$. In each scenario, we form batches of 32 utterances. During inference, we compute global tokens using reference narration utterances, and prepend the text tokens and global tokens before starting the generation, following SparkTTS setup, yielding the following input sequence $[T, G_r]$ where $G_r = (g_{1,r}, \ldots, g_{32,r})$ are global tokens computed from the reference narration.

C.2 Training from Scratch

Here, our experiments include two variants. The first variant is similar to SparkTTS and uses the text to synthesize as the text-condition, with the unchanged sequence $[T, G, S]$ and training objective L. The second variant leverages contextual information around quotations by replacing the text condition with the sequence $T' = [c^l, start, T, end, c^r]$ where $c^l = (c^l_1, \ldots, c^l_n)$ and $c^r = (c^r_1, \ldots, c^r_m)$ are the text tokens from the left and right context respectively, $T = (t_1, \ldots, t_N)$ are the tokens from the text to synthesize, and $start$ and end are special delimiter tokens. Note that we also replace each quotation appearing in c^r and c^l with a special token. Therefore, most of the contextual information can be seen as narrative details. We use a context of one paragraph for both c^r and c^l. We train each variant for 10 epochs with Adam optimizer, a peak learning rate of 3e-5 and a batch size of 32. A cosine schedule is used with 20K warmup steps.

D F5-TTS Experiments

Our experiments also involve fine-tuning F5-TTS on LibriQuote's expressive subset $\mathbf{Q_f}$. We use the official fine-tuning script from the f5-tts repository[9], starting with the official `V1-Base` checkpoint. We use a batch size of 38400 frames per GPU (with

[9] https://github.com/SWivid/F5-TTS/blob/main/src/f5_tts/train/train.py

	Utterance	Verbs			Adverbs			Adjectives			Nouns		
Support	400	344			56			28			14		
	Runtime (s)	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Llama3.3-70b	64	0.82	0.97	0.89	0.65	0.86	0.74	0.64	0.43	0.51	0.80	0.29	0.42
Llama3.3-70b (SC)	-	0.91	0.91	**0.91**	0.8	0.54	0.65	1.0	0.05	0.09	1.0	0.07	0.13
Phi4	6	0.85	0.93	0.89	0.74	0.81	0.77	0.59	0.48	0.53	0.45	0.36	0.32
Phi4 (SC)	-	0.92	0.86	0.89	**0.95**	0.61	0.74	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-32b	19	0.86	0.93	0.89	0.69	0.92	0.79	0.5	0.76	0.6	0.43	0.71	0.54
Qwen3-32b (SC)	-	0.92	0.89	0.9	0.81	0.85	**0.83**	1.0	0.05	0.09	1.0	0.14	0.25
Qwen2.5-72b	62	0.88	0.88	0.88	0.62	0.85	0.71	0.44	0.71	0.55	0.21	0.5	0.21
Qwen2.5-72b (SC)	-	**0.98**	0.8	0.88	0.93	0.73	0.82	0.67	0.19	0.25	0.38	0.21	0.16

Table 6: Narrative information extraction results for all models and information type. Models with # parameters $\geq$ 70b are 4-bit quantized. Runtime is calculated in seconds on a single H100 card, with vLLM as the inference backend. Best results for precision and F1 are indicated in bold and second best results are underlined.

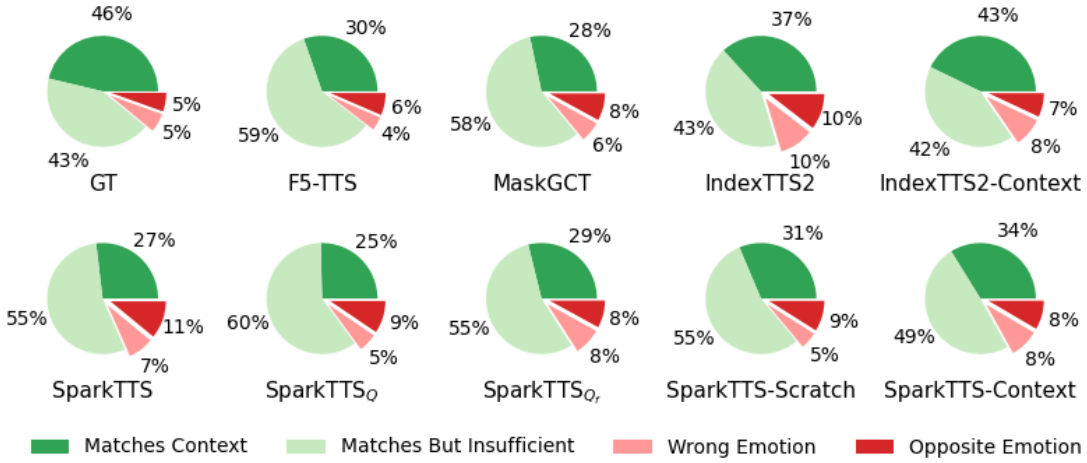

Figure 4: Proportion of failure cases per model predicted by Gemini-2.5 Flash.

2 H100 GPUs) and a maximum of 32 sequences per batch. The learning rate is set to 7.5e-5, and we use 20000 warmup steps, training for a total of 3 epochs.

E Contextual Metrics

With recent breakthroughs in Large Language Models (LLMs) and Large Audio Language Models (LALMs) understanding capabilities, automated evaluations of medium-sized dedicated test samples with LLM-as-a-judge (Gu et al., 2025) or LALM-as-a-judge (Manku et al., 2025) have become increasingly popular. These approaches delegate the usual human inspection of some test samples to LLMs or LALMs to provide an aggregated response, which was often proved to correlate strongly with human judges (Gu et al., 2025; Manku et al., 2025). While human judgements would be ideal to evaluate to which extent a quotation sample convey the appropriate emotion and/or voice intent given a particular book context, subjective experiments at scale are expensive in terms of time and costs, and require dedicated knowledge to construct and analyse. LALM-as-a-judge is thus used as a replacement, noting that a model individual conclusions are often themselves biased, and that only aggregated scores are usually worth exploring.

We follow Manku et al. (2025) and use Gemini-2.5 Pro as the model judge. We hypothesize that the combined ability of LALMs to perform audio and text reasoning is valuable in our particular setup, where text reasoning is suited for understanding the right intent to convey for a particular quotation, and audio reasoning for understanding if a sample appropriately exhibits this intent. We propose two metrics that are built by modifying prompts provided in Manku et al. (2025): ContextMOS and Win-Rate. ContextMOS takes a single audio sample, a quotation text and its context, and prompts Gemini-2.5 Pro to assign a score between 1 and 5, where 1 indicates no alignment between the sample intent and the intents stemming from the context, while 5 indicates complete alignment and natural-

ness in conveying this intent. In contrast, Win-Rate takes a ground-truth audio sample and a synthesized sample, a quotation text and its context, and chose which audio sample conveys the right intent the most faithfully. Note that we allow ties, and define the Win-Rate to of a system T_i relative to the ground-truth as the following:

$$W(T_i) = \frac{\sum(\text{winner} = \text{index}_i) + 0.5 \cdot \sum(\text{winner} = 0)}{n}$$

where 0 indicates ties and n is the number of ratings. An example prompt for ContextMOS is provided in Figure 8 and in Figure 7 for Win-Rate. A similar structure is created for the Win-Rate prompt. All prompts can be found in the dataset repository.

F Baselines

In this section, we describe in more details the baselines used in our experiments.

SparkTTS is an autoregrssive TTS system that builds upon Qwen2.5-0.5M. It also uses BiCodec, a speech codec model that transforms 16kHz sampling rate speech into two types of discrete tokens: global and semantic tokens. Global tokens are derived from representations of an ECAPA-TDNN model fine-tuned for speaker-verification, hence capturing speaker-related information such as timbre. The ECAPA-TDNN representations are mapped to a set of 32 hidden features using learned laten queries, which are then discretized using Finite Scalar Quantization (Mentzer et al., 2023). Semantic tokens are derived from representations of the 11th, 14th and 16th layers of Wav2Vec 2.0 (XLSR-53) (Pasad et al., 2023). These features are then quantized using single-codebook vector quantization, using factorized codes as done in DAC (Kumar et al., 2023a). Tokens are converted back to raw audio using a Convolutional Neural Network (CNN) Decoder composed of ConvNeXt blocks. The language modeling framework is used to learn to generate semantic tokens, conditioned on the text-to-synthesis tokens and global tokens (which are prepended in the autoregressive sequence).

F5-TTS is a non-autoregressive diffusion TTS system based on Flow Matching that does not rely on phonetic alignment nor duration predictor, operating on 24kHz sampling rate speech. During training, following the text-guided speech infilling task, speech mel spectrograms are first masked and then concatenated with text tokens. The resulting sequence is inputted to a Diffusion Transformer (DiT) (Peebles and Xie, 2023), which learns to generate the masked spectrogram region from gaussian noise using Flow Matching. During inference, the reference speech mel spectogram serves as the unmasked region, and is concatenated with the text to synthesize the input sequence. Sway sampling is used along with an Ordinary Differential Equation (ODE) solver to generate mel spectrograms that are converted back using Vocos (Siuzdak, 2024).

MaskGCT is also a non-autoregressive TTS system, that follows a two-step masked generative modeling paradigm. It leverages two speech codec models to produce semantic and acoustinc discrete units from 24kHz sampling rate speech. Semantic tokens are derived from representations of the 17th layer of W2v-BERT 2.0, that are converted to discrete units using a single codebook of size 8,192 via classic Vector Quantization (VQ). Acoustic tokens are produced following RVQGAN (Kumar et al., 2023b), and 8 codebooks of size 8,192 are used via Residual VQ (RVQ). To generate raw audio, MaskGCT follows a two-step process: first, it uses a Transformer to generate semantic tokens by iteratively predicting masked tokens conditioned on the text to synthesize, prompt tokens and previously generated semantic tokens ; second acoustic tokens are generated following a similar iterative procedure, but using an architecture similar to SoundStorm (Borsos et al., 2023), and the generative process is conditioned with the prompt acoustic tokens and previously predicted semantic tokens. Finally, predicted acoustic tokens are converted back to raw audio using Vocos. Additionally, MaskGCT employs a Flow-Matching duration predictor to predict the target speech duration during inference.

IndexTTS2 is a recent autoregressive TTS system that enables fine-grained control over duration and emotions. It incorporates three modules: a Text-to-Semantic (T2S) module, a Semantic-to-Mel (S2M) module and a vocoder. The T2S module autoregressively generates semantic tokens derived from a dedicated Neural Speech Tokenizer, trained with the standard language modeling task, while the S2M module is trained to generate faithful mel-spectrogram from semantic tokens, trained with flow matching. The vocoder converts mel-spectrogram back into raw audio. The core innovation of IndexTTS2 relies on its capacity to control the duration and emotion of synthesized speech. In particular, a soft-instruction module derived from

Qwen3 predicts an emotion distribution from the text description, that is used to condition the speech synthesis. This decoupled approach allows innovative disentanglement between emotional expression and speaker identity – an approach particularly suited for the task of synthesizing character quotation – leading to large improvements in expressive TTS.

G Subjective Experiments

Human subjects for the listening tests were recruited from our organization via voluntary participation. Participants were indicated that they were part of a study involving the comparison of AI-generated speech and human speech. They were also advised to use headphones and to complete the study in a quiet environment.

For the CMOS experiments, we ask the participants the following question: *How expressive is this recording? Please focus on aspects of style (timbre, emotion and prosody), and ignore the aspects of content, grammar, or audio quality*. For each instance, we give participants two audios containing one human sample and one generated sample, in a randomized order. We ask the participants to rate if the first audio sounds more expressive than the second on a $[-3, 3]$ scale.

For the MOS experiments, we ask the participants the following question: *How natural (i.e human-sounding) is this recording? Please focus on examining the audio quality and naturalness, and ignore aspects of style (timbre, emotion, prosody)*. For each instance, we give participants a single audio, and ask them to rate the naturalness of the sample on an 1-5 Likert scale with 0.5 intervals.

We provide screenshots of the annotation platform in Figure 9 and Figure 10.

H Investigating Contextual Matching

We showed that IndexTTS2 is able to produce high-quality expressive speech. However, we found during listening that some samples could exhibit a very strong emotion that is completely different from what one could expect from the context. To investigate the proportion of such cases, we leverage the reasoning traces produced by Gemini-2.5 Pro when providing a score for ContextMOS. These traces contain fine-grained details describing the expected emotion from the context, and what is conveyed appropriately or wrongly in the provided samples, sometimes with precise timestamps. To provide an overview of failure cases, we provide Gemini-2.5 Flash with a reasoning trace (in text format) only and prompt it to categorize each sample based on how its emotion matches the context. Figure 4 shows the result for each model and Figure 5 provides with the prompt used. While IndexTTS2 shows a large number of samples that convey the appropriate emotion, it also shows the largest number of samples with the wrong or opposite emotion. Providing contextual information to IndexTTS2 largely reduces the ratio of wrong/opposite emotions. Similarly, SparkTTS trained from scratch with contextual information that exhibits the largest ratio of samples with appropriate emotions across all variants of SparkTTS.

Overall, these results suggest that both models and human samples seem to insufficiently depict the appropriate emotions for a given quotation, showcasing challenges in faithfully impersonating characters during book reading.

Indeed, LibriVox recordings are recorded by amateur audiobook narrators, thus not all utterances are recorded with the same effort to deliver a perfect reading, which is the reason why ContextMOS ground-truth scores are not considerably higher. We believe that replicating our experiments with professional audiobook narrators samples would likely exhibit higher ContextMOS, drawing a clearer gap between human and synthesized quotations.

I Hardware Information

All experiments done in this paper were conducted on a single compute node equipped with 4 NVIDIA H100 with 80GB of GPU RAM.

You will be given a paragraph that includes reasoning traces of a previous audio analysis. This analysis was performed by a LLM that was tasked to judge wheter a speech sample of an audiobook utterance was appropriately delivered in terms of prosody and emotion, given the book context in which the utterance occurs.
Given the reasoning trace and the assessed score on a scale from 1 to 5, where 1 indicates complete failure at delivering the appropriate emotion, and 5 indicates complete success, your goal is to **classify** the prediction in **4 error categories**:

Error Categories:
1. Success: the speech sample successfully delivered the right prosody and emotion.
2. Opposite: the speech sample delivered natural prosody and emotion, but it was the opposite emotion/prosody that was expected given the context.
3. Insufficient: the speech sample delivered insufficient prosody and emotion given the context.
4. Wrong: the speech sample delivered natural prosody and emotion, but it was not the one expected given the context.

Now, you will be provided the reasoning trace, that include the prediction reasoning in natural language, and the integer score in a 1-5 scale.

Reasoning Trace:
{{{reasoning_trace}}}

output_format
You will output a json dictionary as follows:
{
"error_reasoning": str = Reasoning chain to explain why the prediction should be either one of the **Error Categories**.
"prediction": int = Your prediction between 1 and 4 based on the **Error Categories**.
}

- Note: Ensure the json structure is followed and the json output **MUST** be parsable without errors.(For example, escape the quotes wherever you add them inside a field of the json, all brackets and braces should be correctly paired.)

Figure 5: Example prompt used with Gemini-2.5 Flash for predicting error categories.

Admit	Announce	Argue	Assure
Babble	Bark	Bawl	Beg
Bellow	Bemoan	Blabber	Bleat
Bluster	Boast	Brag	Breathe
Cackle	Chant	Cheer	Chirp
Chirrup	Cluck	Complain	Confide
Cough	Drawl	Exclaim	Falter
Fuss	Giggle	Groan	Grumble
Growl	Grunt	Hiss	Holler
Hoot	Howl	Hum	Implore
Jabber	Jibber	Laugh	Moan
Mouth	Mumble	Murmur	Mutter
Nag	Pant	Pester	Prattle
Pronounce	Ramble	Rebuff	Retort
Roar	Sass	Scream	Screech
Shout	Shriek	Sing	Sigh
Snap	Snarl	Snicker	Sniff
Snigger	Snivel	Sob	Spit
Sputter	Squeak	Squeal	Stutter
Taunt	Tease	Trill	Wail
Weep	Whimper	Whine	Whisper
Whistle	Yell	Yelp	Hesitate
Pause			

Table 7: List of speech verbs used for building the filtered subset $\mathbf{Q}_f$.

You are an expert in linguistic. You like to read books and excel at analyzing dialogues in literature.

Given a small narrative passage, where each quote content is masked, your role is to extract speech verbs, adverbs, adjectives and nouns that indicate how a target quotation is being uttered. You will be given a target quotation marked with [TARGET] that occur in the passage. You need to extract speech verbs, adverbs, adjectives and nouns that follow these criteria:
- It must be either a speech-verb, an adverb, a noun or an adjective.
- It must be one word only.
- It must be a speech descriptor of the target quotation.
- If an adverb, it must be a descriptor of one of the speech-verbs describing the target quotation.
- If a verb, **ensure that it is a speech-verb and not a verb describing anything else than speech.**

Note that multiple speech-verbs can be found and that target quotations can have no associated speech-verbs.

Return a dictionary where keys are the words extracted in the final step and the values is another dictionary with keys 'id' for the paragraph id of the word, 'type' for the word type (verb, adverb, adjective, noun) and 'confidence' an integer between 0 and 10 measuring how confident you are in your prediction, 0 being not confident at all and 10 being sure you are right.
Before creating the dictionary, make sure again that all the criteria above are respected. **Only generate this dictionary and nothing else. Return an empty dictionary if no words were found.**

Passage:
[0] Ella handed the notebook to Jay, eyes uncertain.
[1] Jay flipped through the sketches, pausing at one. [QUOTE_1] She nodded.
[2] [TARGET] whispered Ella slowly.

Target quotation: [TARGET]

Answer:
{ "whispered": { "id": "2", "type": "verb", "confidence": 10 }, "slowly": { "id": "2", "type": "adverb", "confidence": 10 } }

Passage:
[0] She went on, half laughing
[1] [TARGET] Then we went to the park, and he said [QUOTE_1]

Target quotation: [TARGET]

Answer:
{ "went": { "id": "0", "type": "verb", "confidence": 9 }, "laughing": { "id": "0", "type": "verb", "confidence": 9 } }

...

Passage:
[0] I said I had got it on the boat. So then she started for the house, leading me by the hand, and the children tagging after. When we got there she set me down in a split-bottomed chair, and set herself down on a little low stool in front of me, holding both of my hands, and says:
[1] [TARGET]
[2] [QUOTE_1]

Target quotation: [TARGET]

Answer:

Figure 6: Example prompt used with `Phi-4` (with self-reported confidence) to extract narrative information.

Your goal is to judge two audio samples that deliver the same text and analyze if a speech sample seems better than the other one on a particular **evaluation dimension** and determine the winner based on the scoring criterion. You will rate each sample a score between 0 and 3 based on how well the speech corresponding to the target quotation within a particular text called **contextual_text** is delivered, then do their comparative analysis and provide your final judgement. A sample should sound realistic and human-like, and should capture the specific nuances of the text.

You will be provided with the **contextual_text** which is the full text containing the quotation delivered in the speech sample and other side information. the **text_category** and the **evaluation_criterion** corresponding to the **text_category**, in which you will be made aware of the **evaluation dimension** you will focus on, and the **scoring criteria** you will use to score the samples. You will also be provided with the **output_format**, which dictates the format of the output you need to follow as a judger. Finally, you will first be provided with the speech sample 1 **sample_1** and then speech sample 2 **sample_2**.

contextual_text
{{{context}}}

text_category
Emotions
evaluation_criterion
Evaluation Dimension:
- In this category, we want the expression of natural emotions and contextual voice acting, using variations in pitch, loudness, rhythm, etc.
- The sample contain speech delivered from a quotation present in the **contextual_text** and identified with <quote_start> and <quote_end> markers. Delivering a good quality speech means showing natural and strong expressiveness for the quoted dialogue, which aligns with the contextual information.

Example:

{{{Examples}}}

Rating Scale:
1: Fails to express emotions or contextual voice acting (whispering etc..).
2: The rendered emotions or contextual voice acting are not very natural.
3: Synthesises the quotation with natural emotions or contextual voice acting.
Note: - The **contextual_text** will not explicitly state the emotion for the quotation, you have to infer that from the context. - Samples demonstrating exaggerated expressiveness should not be rewarded more **UNLESS** the expressiveness features are relevant to the delivery of appropriated emotion and/or voice acting.

Reasoning guidelines:
{{{reasoning_guidelines}}}

NOTE: If the speech is very poor and does not synthesise the text correctly, you will provide a score of 0. GLOBAL CONSIDERATIONS(**VERY IMPORTANT FOR COMPARISON**): - It is imperative to compare the two samples **ONLY** on the basis of the **evaluation_dimension**, that means, you **WILL NOT** let the following types of **BIASES** affect your judgement:

{{{bias_mitigation}}}

output_format

You will output a json dictionary as follows:
{
"reasoning_sample_1": str = ...
"reasoning_sample_2": str = ...
"comparison": str = ...
"score_1": int = ...
"score_2": int = ...
"winner": int = ...
}
Now you will be provided with the speech from sample 1, please analyze it carefully.

sample_1
{{{audio_1}}}

Now you will be provided with the speech from sample 2, please analyze it carefully. After that provide the judgment following the **output_format** ensuring parsability.

sample_2
{{{audio_2}}}

Figure 7: Example prompt used with Gemini-2.5 Pro for Win-Rate. Provided audios are raw wav files.

Your goal is judge a single audio sample that deliver a specific quotation text. You will be provided with the **contextual_text** which is the full text containing the quotation delivered in the speech sample and other side information. The target quotation that is spoken in the speech sample will be identified with <quote_start> and <quote_end> markers. Given the **contextual_text** and the target quotation, you will rate the sample on a score between 0 and 5, based on how well the corresponding speech sample good quality speech means shows natural and strong expressiveness for the quoted dialogue, which aligns with the contextual information. In particular you will focus on aspects of emotions and voice acting and ignore any other factors. You will also be provided with the **output_format**, which dictates the format of the output you need to follow as a judger. Finally, you will be provided with the speech sample **sample**.

contextual_text
{{{context}}}

evaluation_criterion
- We want to evaluate the expression of natural emotions and contextual voice acting, using variations in pitch, loudness, rhythm, etc.

Example:

{{{examples}}}

Rating Scale:
1: Fails completely to express emotions or contextual voice acting (whispering etc..).
2: Delivers an emotion or contextual voice acting that does not match the contextual information.
3: The rendered emotions or contextual voice acting are not very natural.
4: The rendered emotions or contextual voice acting is matching the contextual information but is not strong enough.
5: The delivered speech contains very natural emotions or contextual voice acting that perfectly align with the contextual information.

NOTE: If the speech is very poor and does not match the text correctly, you will provide a score of 0. GLOBAL CONSIDERATIONS(**VERY IMPORTANT FOR COMPARISON**):
- It is imperative to judge the sample **ONLY** on the basis of the **evaluation_criterion**, that means, you **WILL NOT** let the following types of **BIASES** affect your judgement:
- The acoustical quality of the audio, background noise or clarity.
- The gender and timbre features of the speaker. - Any other factors that are not related to the **evaluation_criterion**.
- Samples demonstrating exaggerated expressiveness should not be rewarded more **UNLESS** those features are relevant to the **evaluation_criterion**

Reasoning guidelines:
{{{reasoning_guidelines}}}

NOTE: If the speech is very poor and does not synthesise the text correctly, you will provide a score of 0. GLOBAL CONSIDERATIONS(**VERY IMPORTANT FOR COMPARISON**): - It is imperative to compare the two samples **ONLY** on the basis of the **evaluation_dimension**, that means, you **WILL NOT** let the following types of **BIASES** affect your judgement:

{{{bias_mitigation}}}

output_format

You will output a json dictionary as follows:
{
"reasoning": str = Reasoning chain based on the **Reasoning guidelines:**.
"score": int = Your score for the speech sample between 0 and 5, based on the **evaluation_criterion** and what you have mentioned in "reasoning".
}
Now you will be provided with the speech sample, please analyze it carefully.

sample
{{{audio}}}

Figure 8: Example prompt used with Gemini-2.5 Pro for ContextMOS. Provided audios are raw wav files.

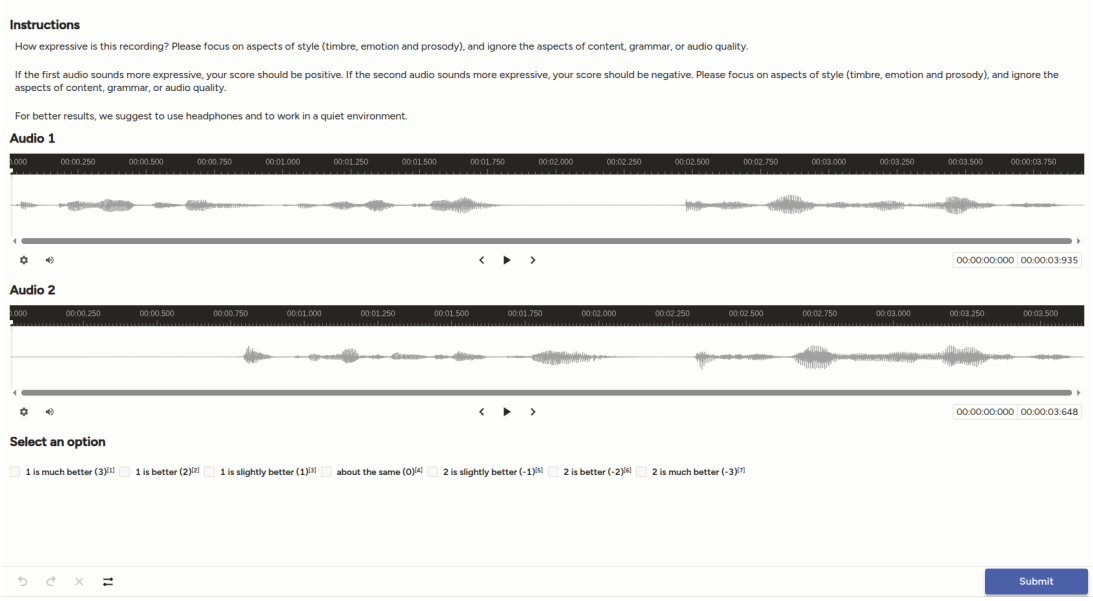

Figure 9: Screenshot of the platform used for the CMOS experiment.

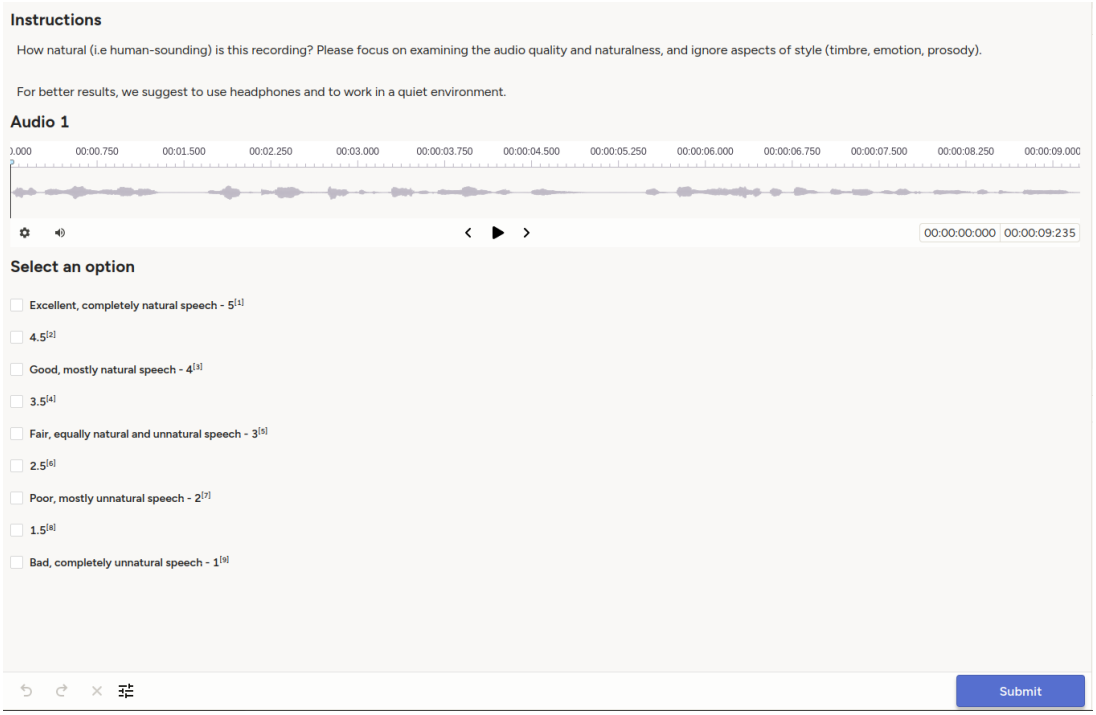

Figure 10: Screenshot of the platform used for the MOS experiment.