A STUDY OF RATERS' SENSITIVITY TO INTER-SENTENCE PAUSE DURATIONS IN AMERICAN ENGLISH SPEECH

Paul Owoicho[*†], *Joshua Camp*[‡], *Tom Kenter*[‡]

[†]University of Glasgow, UK [‡]Google, UK

ABSTRACT

Inter-sentence pauses are the silences that occur between sentences in a paragraph or a dialogue. They are an important aspect of long-form speech prosody, as they can affect the naturalness, intelligibility, and effectiveness of communication. However, the user perception of inter-sentence pauses in long-form speech synthesis is not well understood. Previous work often evaluates pause modelling in conjunction with other prosodic features making it hard to explicitly study how raters perceive differences in inter-sentence pause lengths. In this paper, using multiple text-to-speech (TTS) datasets that cover different content types, domains, and settings, we investigate how sensitive raters are to changes to the durations of inter-sentence pauses in long-form speech by comparing ground truth audio samples with renditions that have manipulated pause durations. This experimental design allows us to draw conclusions regarding the utility that can be expected from similar evaluations when applied to synthesized long-form speech. We find that, using standard evaluation methodologies, raters are not sensitive to variations in pause lengths unless these deviate exceedingly from the norms or expectations of the speech context.

Index Terms— Speech synthesis evaluation, TTS

1. INTRODUCTION

The nature of read and spontaneous speech is such that speakers incorporate contextual pauses to aid the comprehension of the message being conveyed. Inter-sentence pauses in particular can be used to signal a change in topic or tone, draw attention to a key point, create anticipation for what comes next, or project confidence and clarity [1, 2]. As such, pause modelling in long-form speech synthesis has garnered some research attention, given the hypothesis that a fully human-like implementation of contextual inter-sentence pausing - in addition to other temporal aspects of speech such as syllable prolongations and overall timing structure - will lead to more fluent, natural, and intelligible sounding speech [2].

While some of these modelling approaches have led to sophisticated algorithms and techniques, the extent to which end-users appreciate these efforts remains unclear. Previous work often tackles the pause modelling problem in tandem with other prosodic properties of long-form speech - such as rhythm, stress, tone, and intonation - making it difficult to assess the contribution of pause modelling approaches on the naturalness of the synthesised speech [3, 4, 5]. More so, these approaches are often compared against baselines that use blanket inter-sentence pauses (e.g. 200ms) [6, 7]. While intuitive, the results of our experiments in this work challenge the definitiveness of the outcomes of such setups.

In this paper, we seek to understand how end-users perceive inter-sentence pause alterations in long-form speech. Specifically, using a mix of proprietary and publicly-available text-to-speech (TTS) datasets, we ask raters to state their preference between ground truth audio samples (i.e. as recorded by the speaker) and samples that we alter to have pauses whose lengths are manipulated in various ways. Note that apart from the altered *inter-sentence* pauses, the audio samples are identical, allowing us to isolate the effects of the varying the pause lengths. We restrict our focus to these pause types because of their well-studied effects on speech perception in the literature [8, 9, 10, 11]. Moreover, many modern approaches to long-form speech synthesis generate speech sentence by sentence, incorporating blanket pauses in between [12, 13]. Thus, our work aims to provide insights into the significance of contextually appropriate inter-sentence pauses in the quality of long-form synthesised speech, and to bear significance on future work on inter-sentence pause modelling and evaluation.

Our experiments show that, on average, our raters are unable to perceive variations in inter-sentence pause lengths unless substantial deviations from typical pause lengths occur. We hypothesise that this demonstrates a shortcoming in existing subjective TTS evaluation methods for assessing inter-sentence pauses, as they do not appear to capture the subtle effects of speech in this respect. This has a bearing on future work dealing with pause modelling, and, in particular, on the evaluation scenarios involved.

2. RELATED WORK

In this section we discuss work related to two aspects of the current study: the perception of pauses, and the evaluation of speech.

[*]Work done during an internship at Google, UK.

2.1. The Perception of Pauses

Reich [8] finds that the location of pauses within sentences influences the listener's ability to recall the salient parts of the sentence. Lass [9] makes a related observation, noting that intra- and inter-sentence pauses affect the perceptions of oral reading rates. More recently, Fors [10] uncovers the significance of pauses in conversation, finding that pauses matter in conversational turn-taking and turn-yielding. Similarly, Roberts and Francis [14] find that pauses at or beyond 600 milliseconds tend to have communicative meaning in social contexts, claiming that such pauses are considered too long for speech planning and production. Perhaps closest to our work is Smith [11], who finds that listeners prefer read speech with ground truth inter-sentence pauses to read speech with pauses manipulated to be the average duration from the corpus. Smith, however, also manipulates the speaking rate, making the contribution of inter-sentence pauses to this finding unclear.

Unlike our work, the studies mentioned above either investigate both intra- and inter-sentence pauses simultaneously, or modify other time-related parameters of speech alongside inter-sentence pauses. This implies that their findings are not attributable to inter-sentence pauses alone.

2.2. The Limitations of Subjective TTS Evaluation

Chiang et al [15] present an example of the limitations of speech evaluation, pointing to ranking inconsistencies in the results of ten mean opinions score (MOS) evaluations of three TTS models. Specifically, they find that variances in factors such as the qualification and location of raters, instructions provided to raters, and even the choice of crowdsourcing platform all have a bearing on the outcomes of subjective TTS evaluation. In a similar vein, Clark et al [16] find that the presentation of the audio samples influences how they are rated. For example, when a sentence is evaluated on its own without any context, the average rating it receives from raters can significantly differ from the rating it gets when the same sentence is heard along with some context. Thus, while the context itself might not require a rating, it still influences the perception of the sentence. Cambre et al [17] use a novel evaluation approach to assess a variety of synthesized and human voices for long-form synthesis. They conclude that while TTS voices are on par with human voices, no voice is superior to the rest across the dimensions evaluated. The implication of this is that, ultimately, the perceived quality of a TTS system depends on the context in which the system will be used. Unfortunately, these nuances are difficult to express in standard A/B and MOS tests. Recent work such as [18] and [19] demonstrate cases in which systems achieve the same or similar MOS, but are distinguishable when targeted evaluation protocols are used, indicating that traditional modes of evaluation such as MOS or preference tests may not be sensitive enough for certain aspects of speech.

Unlike the work mentioned in this section, we focus on inter-sentence pauses only, and the sensitivity raters display when exposed to pauses of different lengths. The aim is to contribute insights regarding the usefulness of evaluations of TTS systems that model such pauses.

3. EVALUATION SETUP

For each dataset (see §3.4), raters are presented with two versions of the same audio stimulus: one with ground truth pauses and one with manipulated inter-sentence pauses. Raters are asked to state which stimulus they prefer in a forced choice task.

3.1. Evaluation Conditions

Based on empirical investigations, we evaluate the following four conditions in an attempt to mimic real-world practice and to explore sensitivity of raters to differences in pause lengths. Note that the way we manipulate the pauses is dataset-specific, and we account for outliers by excluding sentences with pauses that are greater than 1.5 times the inter-quartile range of all pauses in the source dataset.

1. **Groundtruth vs. Short Pauses**: In this setting we investigate if raters are sensitive to hearing speech with inter-sentence pauses of the the shortest length possible. As 0-length pauses can lead to sudden jolts and artifacts, we define a short pause as being 5ms in length across all datasets[1].

2. **Groundtruth vs. Average Pauses**: In this condition, we consider pauses that are near the mean of all pauses in the dataset. As it is common to concatenate the outputs of sentence-level speech synthesis systems with same-length pauses of a default length (e.g. 200ms), this condition helps us understand how a human's inherent inter-sentence pausing behaviour compares with real world practice.

3. **Groundtruth vs. Long Pauses**: In this setting we investigate if listeners are sensitive to speech with inter-sentence pauses at the upper end of the pause length distribution.

4. **Groundtruth vs. Inverse Pauses**: Here, we replace ground truth pauses that are greater than the dataset-average with a short pause, and pauses shorter than the dataset-average with a long pause (both as defined above). This setup is aimed at being the most disruptive to the raters.

[1] Due to limitations arising from automatic alignment, the final phoneme of the initial sentence and the first phoneme of the follow-on sentence may have some silence aligned to them, meaning that in reality some pauses may be longer than 5ms.

Table 1. Pause values we use for each condition evaluated in our experiments across all datasets in our collection.

Dataset	Short pause (ms)	Average pause (ms)	Long pause (ms)
LibriTTS	5	370	1,000
CALLHOME	5	700	2,500
News Data	5	200	700

With the exception of inverse pauses, the conditions we evaluate feature *blanket pauses*. That is, we concatenate the ground truth speech signals with the same duration of pause between all segments. Table 1 details the pause values we use in our experiments. We highlight that our chosen experimental values are designed to push the boundaries of potential observations.

3.2. Audio Sample Generation

We collect 1,000 ratings for each evaluation condition across our collection, by generating at most 200 samples from each dataset. In an effort to mitigate rater fatigue, we make sure each sample features speech by only one speaker spanning 3 to 5 sentences. We use the one speaker condition to stay to close to the experience of listening to an actual speech synthesis system.

3.3. The Style Preference Question

For each comparison pair, we ask raters to choose the sample they prefer as a style of speech, where the style of speech is based on the dataset the samples are derived from. Thus, for CALLHOME, for example, we ask, *Which side sounds better as a telephone conversation?*. For LibriTTS and News Data, the intended styles are "an audiobook narrator" and "a news reader reading the start of a news article" respectively. We believe this gauges how well each sample fits the implicit human expectation of speech in the target style and context, while not explicitly asking about the pauses in order to avoid bias.

3.4. Data

We utilise a diverse range of proprietary and publicly-available spontaneous and read speech datasets to ensure comprehensive coverage across various speaking styles, content types, domains, and settings. In addition, each dataset contains recordings from multiple speakers, enhancing the breadth of our collection.

3.4.1. LibriTTS

LibriTTS is a large-scale multi-speaker corpus of English speech and text that contains 585 hours of speech from 2,456 speakers, covering various topics and domains [20]. It is derived from the LibriSpeech dataset [21], a collection of audiobooks from LibriVox [22] and Project Gutenberg [23]. For average pauses we use the dataset mean of 370ms, and 1 second for long pauses (90th percentile). Utterance pairs with 0ms pauses are filtered as they likely indicate boundary alignment errors. We note that there is 200ms of padding on either side of utterance boundaries, except for short pauses where this padding is stripped.

3.4.2. CALLHOME American English Speech

The CALLHOME American English Speech dataset contains 120 spontaneous telephone conversations between native speakers of English, recorded by the Linguistic Data Consortium [24]. The conversations cover a variety of topics and domains, from family and friends to hobbies and travel. For average and long pauses, we use 700ms (close to the dataset average of 680ms) and 2,500 ms (the 99.5th percentile pause) respectively. Because the phone conversations are noisy, naively expanding the ground truth pause segments to the desired length results in samples with audible cuts. To mitigate this, we expand the pause segment with randomly selected sub-segments to reach the desired pause length.

3.4.3. News Data

We also experiment with a proprietary dataset of read news articles. The dataset consists of 8 speakers reading news articles in an informative style. We focus on the beginning of each article, containing the title, subtitle/author, and first sentence, as these tend to have the most varied inter-sentence pauses. For average pauses, we use 200ms, near the dataset average of 190ms. For long pauses, we use 700ms, the 99.5th percentile pause length. As this dataset has less varied pauses than others, we choose a very long pause length. Due to the small dataset size, we keep 0ms pauses unchanged with no manipulation or filtering.

4. RESULTS AND ANALYSIS

The results of our experiments are shown in Table 2. As can be observed from the table, the preference ratings for ground truth vs. each condition of manipulated pause are close to chance (i.e. 50%) in most cases, suggesting that our listeners generally did not exhibit a strong preference between the audio samples that we present to them. Below, we discuss some observations and key takeways.

Table 2. Results of side by side preference tests comparing ground truth pauses vs manipulated pauses across datasets. Values are the percent of ratings expressing a preference for ground truth pauses with 99% confidence intervals. Bold indicates significance at a p-value of 0.01.

Dataset	vs. Short (%)	vs. Average (%)	vs. Long (%)	vs. Inverse (%)
LibriTTS	53.3 ± 4.09	48.9 ± 4.12	$\mathbf{54.5 \pm 4.08}$	$\mathbf{55.5 \pm 4.06}$
News Data	51.7 ± 4.10	47.9 ± 4.12	$\mathbf{56.6 \pm 4.05}$	$\mathbf{54.2 \pm 4.08}$
CALLHOME	50.6 ± 4.11	53.7 ± 4.09	$\mathbf{74.7 \pm 3.47}$	$\mathbf{54.8 \pm 4.07}$

4.1. The Blindspot for Inter-Sentence Pauses

We find that our listeners, on average do not perceive a difference between ground truth and short pauses, as evidenced by the non-statistically significant 51% - 53% preference across all datasets. This implies that very short and barely noticeable inter-sentence pauses do not negatively impact the perceived quality of speech. Similarly, with preferences close to 50% across all datasets, our listeners do not prefer groundtruth over average pause lengths, which are typically used in practice. Importantly, this indicates that using a fixed pause length of reasonable length between sentences produces speech comparable in quality to contextually-dependent ground truth pauses under our evaluation methodology.

However, raters show a preference for ground truth compared to long pauses deviating far from norm. For example, for the CALLHOME dataset, our listeners prefer groundtruth 74% of the time over samples with long 2,500ms pauses, a statistically significant difference. This suggests that there are speech-dependent thresholds where unnaturally long pauses reduce speech quality. In the case of inverse pauses, that swap short and long durations, we see a preference for groundtruth, but notice that this is only a modest preference (54–56%).

4.2. General Issues in Longform Evaluation

We believe our results point to issues in the evaluation of long-form synthetic speech. While it is possible to find individual instances in which blanket average pauses sound less natural than ground truth pauses, these cases are rare. Therefore, they have little impact on the overall test score. As such, we believe the evaluation of some aspects of long-form speech have a **sparsity problem**, whereby evaluations on random samples of test material are unlikely to uncover real issues. Alternatively, practitioners could consider constructing test sets consisting only of types of inputs that are known to be problematic.

Further, existing evaluation methods might not be sufficient to detect issues real users would notice. In short clips of 3–5 sentences, distinguishing blanket pauses from the ground truth can be challenging. However, this discrepancy might become more apparent to users during longer listening sessions. Therefore, **ecological validity** concerns could be more pronounced in long-form contexts.

4.3. Implications for Future Work

Our results suggest that listeners are insensitive to minor variations in inter-sentence pause durations, as evidenced by the lack of preference for ground truth over short or average pauses. It is only when pauses substantially and consistently deviate from norms and expectations of the speech context that quality judgments are impacted. This highlights potential limitations of standard subjective evaluation methods for assessing pause modeling, as degradations or improvements are likely to go unnoticed. More targeted test set design considering sparsity of inappropriate pauses in natural speech may be needed. Furthermore, our results raise the question of whether existing methodologies can effectively capture nuanced aspects of quality that impact real users as the listening conditions of current tests may not sufficiently replicate real environments for long form speech.

5. CONCLUSION

In this work, we investigate listener sensitivity to inter-sentence pause variations in diverse speech datasets. We find that listeners do not perceive quality differences between groundtruth and manipulated pauses, unless pauses deviate considerably from the dataset norm.

We believe these results have the following implications for future work on long-form speech synthesis: (1) Inter-sentence pauses alone may not significantly impact overall long-form prosody quality, compared to other factors like intonation and rhythm. (2) Carefully designed test sets and evaluation methods may be needed to properly assess pause modeling, since inappropriate pauses are sparse in natural speech. Standard random test samples are unlikely to contain enough cases to show differences. (3) If optimizing inter-sentence pausing, evaluation should consider real usage contexts and listening environments to determine if quality gains are noticeable to end users.

Overall, this work highlights the difficulty of evaluating the effects of subtle temporal aspects of speech such as pausing. Future work should explore more targeted and comprehensive evaluation strategies to better understand the role of inter-sentence pauses in improving long-form synthesis.

6. REFERENCES

[1] Sherry R Rochester, "The significance of pauses in spontaneous speech," *Journal of Psycholinguistic Research*, 1973.

[2] Brigitte Zellner, "Pauses and the temporal structure of speech," in *Fundamentals of speech synthesis and speech recognition*. John Wiley, 1994.

[3] Liumeng Xue, Frank K Soong, Shaofei Zhang, and Lei Xie, "ParaTTS: Learning linguistic and prosodic cross-sentence information in paragraph-based TTS," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[4] Peter Makarov, Ammar Abbas, Mateusz Lajszczak, Arnaud Joly, Sri Karlapati, Alexis Moinet, Thomas Drugman, and Penny Karanasou, "Simple and effective multi-sentence TTS with expressive and coherent prosody," *arXiv preprint arXiv:2206.14643*, 2022.

[5] Shinnosuke Takamichi, Daisuke Saito, Hiroshi Saruwatari, and Nobuaki Minematsu, "The UTokyo Speech Synthesis System for Blizzard Challenge 2017," *The Blizzard Challenge 2017*, 2017.

[6] Alok Parlikar and Alan W Black, "Modeling pause-duration for style-specific speech synthesis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[7] Norbert Braunschweiler and Langzhou Chen, "Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS," in *Eighth ISCA workshop on speech synthesis*, 2013.

[8] Shuli S Reich, "Significance of pauses for speech perception," *Journal of Psycholinguistic Research*, 1980.

[9] Norman J Lass, "The significance of intra-and inter-sentence pause times in perceptual judgments of oral reading rate," *Journal of speech and Hearing Research*, 1970.

[10] Kristina Lundholm Fors, *Production and perception of pauses in speech*, Ph.D. thesis, Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg, 2015.

[11] Caroline L. Smith, "Topic transitions and durational prosody in reading aloud: production and modeling," *Speech Communication*, 2004.

[12] Jingdong Li, Hui Zhang, Rui Liu, Xueliang Zhang, and Feilong Bao, "End-to-end Mongolian text-to-speech system," in *2018 11th international symposium on chinese spoken language processing (ISCSLP)*, 2018.

[13] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[14] Felicia Roberts and Alexander L Francis, "Identifying a temporal threshold of tolerance for silent gaps after requests," *The Journal of the Acoustical Society of America*, 2013.

[15] Cheng-Han Chiang, Wei-Ping Huang, and Hung-yi Lee, "Why we should report the details in subjective evaluation of TTS more rigorously," *arXiv preprint arXiv:2306.02044*, 2023.

[16] Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," in *10th ISCA Speech Synthesis Workshop (SSW10)*, 2019.

[17] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye, "Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[18] Ayushi Pandey, Jens Edlund, Sébastien Le Maguer, and Naomi Harte, "Listener sensitivity to deviating obstruents in WaveNet," in *Proc. INTERSPEECH 2023*, 2023.

[19] Harm Lameris, Joakim Gustafson, and Éva Székely, "Beyond Style: Synthesizing Speech with Pragmatic Functions," in *Proc. INTERSPEECH 2023*, 2023.

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

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015.

[22] Jodi Kearns, "LibriVox: Free public domain audiobooks," *Reference Reviews*, 2014.

[23] Bryan Stroube, "Literary freedom: Project gutenberg," *XRDS: Crossroads, The ACM Magazine for Students*, 2003.

[24] Alexandra Canavan, David Graff, and George Zipperlen, "Callhome american english speech," *Linguistic Data Consortium*, 1997.