Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

# A Study of the Sensitivity of Subjective Listening Tests to Inter-sentence Pause Durations in English Speech

*Paul Owoicho*[1], *Joshua Camp*[2], *Tom Kenter*[2]

[1]University of Edinburgh
[2]Google, UK

paul.owoicho@ed.ac.uk, joshcamp@google.com, tomkenter@google.com

## Abstract

Inter-sentence pauses are silences occurring between sentences in a paragraph or dialogue. They are an important aspect of long-form speech prosody, as they can affect the naturalness and effectiveness of communication. When evaluating the output of long-form speech synthesis systems, it is crucial to understand the sensitivity of commonly used tests to variations in inter-sentence pause durations, as this sensitivity impacts the usefulness of such evaluations. However, 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 differences in inter-sentence pause lengths are perceived. To fill this gap, we investigate the sensitivity of subjective listening tests 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. Using multiple datasets to cover a variety of domains, we find that listening tests are not sensitive to variations in pause lengths unless these deviate from the norm exceedingly. Our evaluation experiments in this study can be considered preliminary work, the findings of which will have implications for evaluation experiments run on actual synthesized long-form speech.

**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 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, 3].

While some of the approaches modelling inter-sentence pauses 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 [4, 5, 6]. More so, these approaches are often compared against baselines that use blanket inter-sentence pauses (e.g. 200ms) [7, 8, 9, 10]. 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 altering inter-sentence pause durations in long-form speech affects the results of subjective listening tests. We restrict our investigation to listening tests designed to evaluate text-to-speech systems, and as such we do not attempt to make psycholinguistic claims about listener perception of inter-sentence pause durations generally. Rather, we are interested in the degree to which subjective listening tests are sensitive to inter-sentence pause durations, to understand whether explicitly modelling inter-sentence pauses can be expected to lead to improved perceived quality the way it is currently measured, or if other evaluation protocols must be developed. Using a mix of proprietary and publicly-available datasets of spontaneous and read speech, we ask raters to state their preference between ground truth audio samples (i.e. as recorded by the speaker) and the same samples that have manipulated inter-sentence pause lengths. Note that apart from the altered *inter-sentence* pauses, the audio samples are identical, allowing us to isolate the effects of varying the pause lengths. We focus on these pause types because of their well-studied effects on speech perception in the literature [11, 12, 13, 14].

As our experiments involve raters assessing recorded speech with manipulated inter-sentence pause lengths, the work presented in this study can be regarded as preliminary. The insights gained from this research will contribute to understanding the potential utility of similar evaluations when applied to actual synthesized long-form speech.

Our experiments show that, on average, speech samples with manipulated inter-sentence pause lengths are not perceived as less appropriate unless substantial deviations from ground truth pause lengths occur. This finding has a bearing on where to allocate future modelling efforts, as the effect of any inter-sentence pause modelling efforts might go unnoticed in subjective evaluations. We hypothesize that our results are due to two primary factors affecting speech synthesis evaluation: **sparsity** and **limited ecological validity**. First, if the phenomena of interest occur or affect perceived quality infrequently, they are unlikely to be detected using test sets constructed using simple random sampling. This issue is exacerbated by the fact that simply adapting traditional evaluation setups to longer speech samples requires much more content to be rated for the same amount of statistical power, meaning that simply increasing test set size or number of ratings may not be practical. Second, ecological validity – the extent to which results obtained in lab studies are applicable in ”real world” contexts – may be worse in long-form. This could be due to the subtlety of the phenomena being investigated (intelligibility is not a concern) or stimulus length; it is a less engaging task and real users engaged in longer listening sessions may notice issues that don’t show up in listening tests.


10.21437/SpeechProsody.2024-94

## 2. Related work

We discuss work related to two aspects of the current study: the perception of pauses, and the evaluation of speech.

### 2.1. The Perception of Pauses

Reich [11] finds that the location of pauses within sentences influences the listener's ability to recall the salient parts of the sentence. Lass [12] makes a related observation, noting that intra- and inter-sentence pauses affect the perception of oral reading rates. More recently, Fors [13] uncovers the significance of pauses in conversation, finding that pauses matter in conversational turn-taking and turn-yielding. Similarly, Roberts and Francis [15] 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 [14], 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 [16] 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 [17] 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 [18] 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 [19, 20] demonstrates 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 of comparative listening tests when used to compare stimuli that differ only in terms of pause length. The aim is to contribute insights regarding the usefulness of such evaluations when applied to TTS systems that model these pauses.

## 3. Evaluation Setup

To allow for the possibility that inter-sentence pause lengths are perceived differently depending on the type of speech, we use multiple datasets in our experiments, covering a variety of styles such as news, audio books (LibriTTS) and telephone conversations (CALLHOME). See §3.2 for more details. For each dataset, raters are presented with two versions of the same audio stimulus in random order: 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

We evaluate the following four conditions.

1. **Groundtruth vs. Short Pauses**: In this setting we investigate inter-sentence pauses at the low end of the pause length distribution. As 0-length pauses can lead to sudden jolts and artifacts, we define a short pause as being 5ms in length across all datasets.
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 noticeable to listeners.

With the exception of inverse pauses, all conditions feature *blanket pauses*. I.e., we concatenate the ground truth speech with pauses of identical length between all segments.

We note that the final phoneme of the initial sentence and the first phoneme of the follow-on sentence may have a tiny portion of silence aligned to them, so some pauses may end up being slightly longer than specified above. As this is only a matter of a small number of frames, it is unlikely to affect any results.

#### *3.1.1. Long pauses*

As noted above, we allow for the possibility of pause lengths being perceived differently depending on the type of speech they occur in. Therefore, we use different values across different datasets because each dataset has its own distribution of pause lengths. For example, a 100ms pause in a conversation might be perceived differently from a pause with the same length in a read news story. We explicitly aim to push the extremes of these pause lengths, both short and long, in an effort to determine the sensitivity at extreme ends of the pause distribution.

Given our corpora, we choose values that balance our attempt to substantively deviate from typical durations while still remaining within the scope of natural speech. We define long pauses as pauses greater than 99.5% of pauses in a dataset, except when specified differently (see §3.2).

Table 1: *Pause values used 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 |

#### *3.1.2. 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 from. For CALLHOME, the raters are asked *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. The difference in questions is intended to gauge 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.

#### *3.1.3. Audio Sample Generation*

We collect 1,000 ratings for each evaluation condition by obtaining at most 200 samples from each dataset. Each sample features speech from one speaker spanning 3 to 5 sentences. We use 3 to 5 sentences in an attempt to strike a balance between rater fatigue and accurately replicating the experience of an extended listening session, which would typically be far longer particularly for news articles and audiobooks.

### 3.2. Data

We utilise publicly available spontaneous and read speech datasets, plus a proprietary news dataset, to ensure comprehensive coverage across various speaking styles and domains. Each dataset contains recordings from multiple speakers. As the datasets are generated at different times, by different entities, differences occur between them that have an impact on our experiments. Table 1 contains relevant descriptive properties of each dataset. We describe the details of each dataset below.

#### *3.2.1. 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 [21]. It is derived from the LibriSpeech dataset [22], a collection of audiobooks from LibriVox [23] and Project Gutenberg [24]. The average inter-sentence pause length is 370ms. We use 1 second for long pauses, which is the 90th percentile. We deviate from the 99.5th percentile used for other datasets as this leads to prohibitively artificial results due to outliers. Utterance pairs with 0ms pauses are filtered as they likely indicate boundary alignment errors (i.e. sentences might be split halfway through, due to tokenization errors). Lastly, only speech samples that are less than 30 seconds long overall are kept to keep the rate of inter-sentence pauses high relative to the rest of the speech.

#### *3.2.2. 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 [25]. The conversations cover a variety of topics, 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. We account for outliers by excluding sentences with pauses that are greater than 1.5 times the interquartile range of all pauses in the source dataset.

#### *3.2.3. News Data*

This proprietary dataset consists of 1044 news articles (20082 sentences, 9680 paragraphs, 19 sentences per article on average), read by 8 speakers in an informative news style. We focus on the beginning of each article, containing the title, subtitle/author, and first sentence, as this is where most inter-sentence pause variation is. For average pauses, we use 200ms, near the dataset average of 190ms. For long pauses, we use 700ms, the 99.5th percentile pause length. This dataset also contains up to 200ms of silence aligned to either side of the inter-sentence pause silence, which is stripped in both the ground truth and experimental conditions. As this dataset is smaller than LibriTTS, we do not apply the same filtering: we keep 0ms pauses unchanged with no manipulation, and utterances over 30 seconds are not filtered.

### 3.3. Ratings

As described above, each sample contains 3 to 5 sentences, i.e., 2 to 4 inter-sentence pauses. In the LibriTTS and News Data conditions, 200 samples are each rated by 5 raters, with each rater rating a maximum of 10 samples. In the CALLHOME condition, 100 samples are each rated by 10 raters, again with each rater rating a maximum of 10 samples. An average of 109.25 raters participate in each condition (a minimum of 105 and a maximum of 115). Each rater only rates one condition per dataset and are asked only to participate if they are using headphones in a quiet environment.

## 4. Results and Analysis

The results of our experiments are shown in Table 2.

### 4.1. Results

As can be observed from Table 2, the preference ratings for ground truth versus each condition of manipulated pause are close to chance (i.e. 50%) in most cases, suggesting that listeners generally did not exhibit a strong preference between the audio samples that we present to them.

Interestingly, changing all inter-sentence pauses to short pauses, at the very low end of the pause distribution, does not have a significant impact on how raters rate the speech compared to the original pauses, while providing long pauses con-

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 of a binomial test with the null hypothesis that preference = 50% at a p-value of 0.01.*

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

sistently does. This is surprising, as one would expect that having only short pauses (i.e. virtually not stopping at all between sentences) would be perceived of as a big and noticeable difference. One potential reason might be that the difference in length between the short and average pauses is less than the difference between the average and the long pause. This is in line with the observation that the results for the long pauses in CALLHOME are the most outspoken, which might be explained by the difference between average and long pause length being by far the biggest for this dataset.

The inverse condition — where pauses above average length are swapped for short ones, and pauses with a length below the average are swapped for long ones — is designed to be the most disruptive setting, and indeed, we observe that all experiments are statistically significant. Surprisingly, the outlier effect in the CALLHOME case is not repeated this time. There are few possible explanations for this finding. The results for the short and long conditions suggest that long pauses are disliked and short pauses are tolerated regardless of context, so it may be that replacing a longer-than-average pause with a short pause is preferable to replacing it with an even longer pause. It may also simply be that a mix of pause lengths is preferred when some pauses are very long.

In short, in our experiments, we only observe statistically significant results if the lengths are very far from the mean (in the long pause setting) or are the opposite of what is natural (in the inverse setting).

### 4.2. Long-form Evaluation

Our results point to issues in the evaluation of long-form synthetic speech more generally. While it would be possible to find cases in which blanket average pauses sound less natural than ground truth pauses, these cases are rare and, as such, have little impact on the overall test score. As a result, we believe the evaluation of some aspects of long-form speech suffer from a **sparsity problem**, whereby evaluations on random samples of test material are unlikely to uncover genuine issues. As an alternative, practitioners could consider constructing test sets consisting only of types of inputs that are known to be problematic or to highlight differences between systems. A related issue is the relative sparsity of ratings obtained in multi-sentence evaluations. If we only obtain one rating for every five sentences, we need to get five times as much content rated to achieve the same statistical power as a sentence-level evaluation. Future work could investigate setups in which listeners are presented with longer samples (thereby providing the context necessary to give accurate ratings), but provide multiple data points per sample (thus increasing statistical power).

Lastly, 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 may be challenging. However, this discrepancy might become more apparent during longer listening sessions. Therefore, concerns around **ecological validity** could be more pronounced in long-form contexts. Further research into user-focused evaluation paradigms [26] could help alleviate this issue.

### 4.3. Weber's Law

Our results can be further analyzed through the lens of Weber's law [27] which states that the just noticeable difference (JND) between two stimuli is a constant proportion of their magnitude, the Weber fraction. Applied to inter-sentence pause lengths, if Weber's law holds with a fraction of 0.2, then for a 500ms pause, a pause would need to change by at least 100ms (0.2 * 500ms) to be consciously detected. Our results indicate that a Weber fraction likely exists for pause perception, but may differ across speech contexts. For example, the Weber Fraction for conversational speech pauses may be different from the one for audiobooks. This could explain why only very long pauses lead to raters indicating a difference for CALLHOME dataset. Understanding Weber Fractions would allow more accurate application of perceptual models to predict when pause variations will be noticeable (see, e.g. [28]). Our results indicate that the fractions may be higher than traditionally assumed.

## 5. Conclusion

In this work, we investigate sensitivity if listening tests to inter-sentence pause variations in diverse speech datasets. We find that raters do not perceive quality differences between ground truth 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 speech synthesis.

## 6. Acknowledgements

We would like to express our gratitude to Berkay İnan, Chun-an Chan, Nicolas Serrano, Rob Clark, Ron Hassan, Toby Hawker, Vincent Wan and the wider TTS team at Google for their invaluable insights and contributions at every stage of this project.

# 7. References


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

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

[3] J. KAHNG, "The effect of pause location on perceived fluency," *Applied Psycholinguistics*, vol. 39, no. 3, p. 569–591, 2018.

[4] L. Xue, F. K. Soong, S. Zhang, and L. Xie, "ParaTTS: Learning linguistic and prosodic cross-sentence information in paragraph-based TTS," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[5] P. Makarov, A. Abbas, M. Łajszczak, A. Joly, S. Karlapati, A. Moinet, T. Drugman, and P. Karanasou, "Simple and effective multi-sentence tts with expressive and coherent prosody," in *Proceedings of Interspeech 2022*, 2022.

[6] S. Takamichi, D. Saito, H. Saruwatari, and N. Minematsu, "The UTokyo Speech Synthesis System for Blizzard Challenge 2017," *The Blizzard Challenge 2017*, 2017.

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

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

[9] J. Li, H. Zhang, R. Liu, X. Zhang, and F. Bao, "End-to-end Mongolian text-to-speech system," in *2018 11th international symposium on chinese spoken language processing (ISCSLP)*, 2018.

[10] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

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

[12] N. J. Lass, "The significance of intra-and intersentence pause times in perceptual judgments of oral reading rate," *Journal of speech and Hearing Research*, 1970.

[13] K. Lundholm Fors, "Production and perception of pauses in speech," Ph.D. dissertation, Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg, 2015.

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

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

[16] C.-H. Chiang, W.-P. Huang, and H.-y. Lee, "Why we should report the details in subjective evaluation of TTS more rigorously," *arXiv preprint arXiv:2306.02044*, 2023.

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

[18] J. Cambre, J. Colnago, J. Maddock, J. Tsai, and J. 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.

[19] A. Pandey, J. Edlund, S. Le Maguer, and N. Harte, "Listener sensitivity to deviating obstruents in WaveNet," in *Proc. INTERSPEECH 2023*, 2023.

[20] H. Lameris, J. Gustafson, and Éva Székely, "Beyond Style: Synthesizing Speech with Pragmatic Functions," in *Proc. INTERSPEECH 2023*, 2023.

[21] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proceedings of Interspeech 2019*, 2019.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015.

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

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

[25] A. Canavan, D. Graff, and G. Zipperlen, "Callhome american english speech," *Linguistic Data Consortium*, 1997.

[26] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Éva Székely, C. Tånnander, and J. Voße, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 105–110.

[27] D. Laming, "Weber's law," in *Insid. Psychol. a Sci. over 50 years*. Oxford University Press, 2009.

[28] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," in *INTERSPEECH 2020*, 2020.