Prosody Analysis of Audiobooks

Charuta Pethe
Department of Computer Science
Stony Brook University
Stony Brook, NY, USA
cpethe@cs.stonybrook.edu

Bach Pham
Department of Computer Science
Earlham College
Richmond, IN, USA
bqpham24@earlham.edu

Felix D Childress
Department of Computer Science
Earlham College
Richmond, IN, USA
fdchild22@earlham.edu

Yunting Yin
Department of Computer Science
Earlham College
Richmond, IN, USA
yinyu@earlham.edu

Steven Skiena
Department of Computer Science
Stony Brook University
Stony Brook, NY, USA
skiena@cs.stonybrook.edu

***Abstract*—Recent advances in text-to-speech have made it possible to generate natural-sounding audio from text. However, audiobook narrations involve dramatic vocalizations and intonations by the reader, with greater reliance on emotions, dialogues, and descriptions in the narrative. Using our dataset of 93 aligned book-audiobook pairs, we present improved models to predict prosody (pitch, volume, and rate of speech) from narrative text using language modeling. Our predicted prosody attributes correlate much better with human audiobook readings than results from a state-of-the-art commercial TTS system: our predicted pitch shows a higher correlation with human reading for 22 out of 24 books in the test set, while our predicted volume attribute proves more similar to human reading for 23 out of the 24 books. Finally, we present a human evaluation study to quantify the extent that people prefer prosody-enhanced audiobook readings over default commercial text-to-speech systems.**

***Index Terms*—prosody attribute prediction, text to speech, character embedding**

I. Introduction

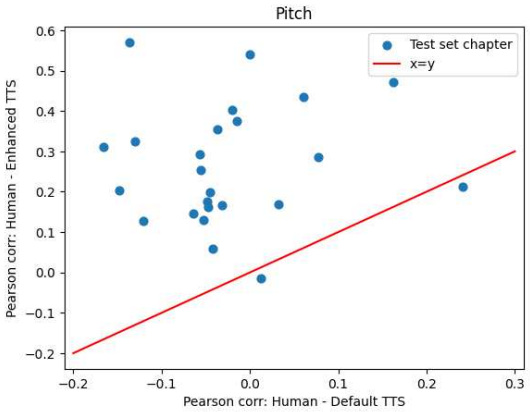
Audio books are actor-narrated recordings of written texts, typically full length novels. Audio books have grown rapidly in popularity in recent years, with annual sales in the United States reaching $1.3 billion dollars in 2020.[1]. The National Library Service for the Blind has recorded tens of thousands of talking books for the sight impaired since 1933, read by an extensive network of volunteers.

In this paper, we analyze audio books from an NLP perspective, with two distinct objectives in mind. First, we study how higher-order NLP analysis of novels (e.g. character identification, quote/dialog analysis, and narrative flow analysis) might be used to build better TTS systems. Second, we are interested as audio books as a source of *annotations* for narrative texts. One of the challenges of NLP on book-length documents is the cost of annotation: reading a novel is a 10-15 hour commitment, making it cost-prohibitive to do human annotation for special-purpose tasks on a large corpora of books. We see human-recorded audio books as a potential solution here: each human recording implicitly contains information about the contents of the texts, which we can programmatically extract by analysing the audio. Our main contributions in this work include:

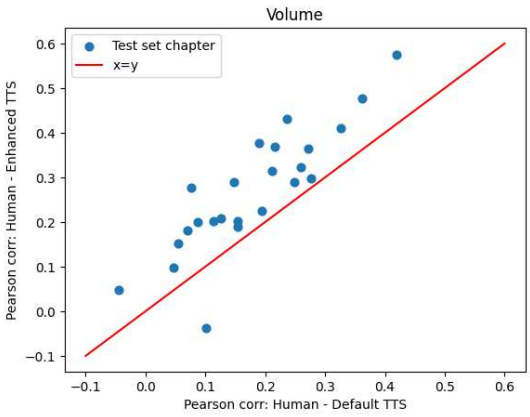
- **Character-level analysis of audiobook reader behaviors** — We quantify the extent to which the gender properties of characters are reflected in audiobook readings. Specifically, in 21 of 31 books where the two lead characters differ in gender, readers used lower pitch and higher volume to portray the male character. Further, readers generally use lower pitch in narrative regions rather than dialog, independent of gender.
- **Models for audiobook prosody prediction** — We address the task of predicting prosody attributes given a narrative text, training a variety of models on audiobook readings to predict prosody attributes (pitch, volume, and rate). The dot plot of Figure 1 presents the human-TTS correlation for our system using a LSTM architecture built upon MPNet embeddings, and a state-of-the-art commercial TTS system (Google Cloud Text-to-Speech) on a test set of the first chapters of 24 books. Our predicted pitch attribute shows a higher correlation with human reading for 22 out of the 24 books, while our predicted volume attribute proves more similar to human reading for 23 out of the 24 books.
- **Human evaluation study for text-to-speech audiobook reading** — We conduct a human evaluation study to understand whether humans prefer to hear our more expressive text-to-speech audio enhanced with Speech Synthesis Markup Language (SSML) over the default commercial Google Cloud Text-to-Speech, a strong baseline. Results are inconclusive, but a small majority (12 of 22) subjects preferred our SSML-enhanced readings.

Over the course of this study, we have developed a dataset of sentence-aligned books and audiobooks comprising 1806 chapters across 93 novels, with their corresponding human-read audiobooks and sentence-level alignment between the text and audio. This dataset has been made publicly available at

[1]https://www.statista.com/topics/3296/audiobooks/



(a) Pitch



(b) Volume

Fig. 1: Enhancing TTS with our predicted prosody attributes show better correlations with human reading than unaided Google Cloud Text-to-Speech.

https://github.com/sbu-dsl/prosody-analysis-of-audiobooks.

This paper is organized as follows. Section II presents related work associated with NLP for books and an overview of the Speech Synthesis Markup Language (SSML) we use in our experiments. Section III describes our methods for the critical phase of alignment between the textual representation of novels and associated audio book recordings. Section IV presents our analysis of reader behavior, quantifying the degree to which voices change with a character's gender. We develop prosody prediction models from text in Section V, culminating in a human preference study reported in Section VI. Future directions for our work are discussed in Section VII.

II. Related Work

A. NLP for books:

Recent work focused on performing NLP tasks on novels [1]–[3] has facilitated the analysis and downstream use of book artifacts [4]–[8]. These systems employ CoreNLP [9], Stanza [10], and SpaCy [11] for NLP annotation tasks. Further, methods to generate character embeddings and analyze relationships between characters [12]–[15] provide additional information about characters apart from their mentions in the book.

```
<speak>
<prosody pitch="-3.19%" volume="-0.11dB" rate="88.83%">
Somehow I felt glad that Jonathan was not on the sea last
night, but on land.</prosody>
<prosody pitch="+26.69%" volume="+1.79dB" rate="136.09%">
But, oh, is he on land or sea?</prosody>
<prosody pitch="+38.79%" volume="+0.94dB" rate="116.57%">
Where is he, and how?</prosody>
</speak>
```

Fig. 2: Sample SSML text with custom prosody attributes.

B. SSML Overview

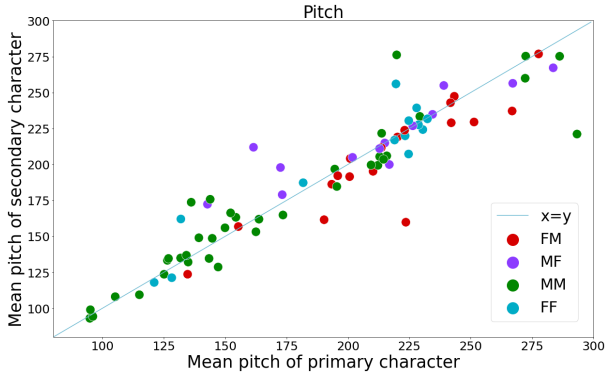
Speech Synthesis Markup Language (SSML) [17] is a markup language that provides a standard way to mark up text for the generation of synthetic speech, allowing for more fine-grained control over the generated audio. In this work, we are specifically interested in the `<prosody>` tag, which can be used to control three different attributes: pitch, volume, and rate of speech. We explore the role of audio in providing additional context and emotional cues [18]. An example of SSML is shown in Figure 2.

III. Audio/Text Alignment Methods

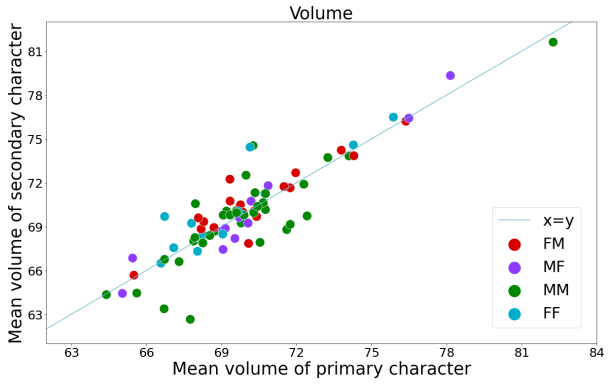
We present a dataset of pairs of books and their corresponding audiobooks from the Project Gutenberg dataset, along with the extracted prosody data from the human-read audiobooks. The dataset contains a total of 1806 chapters across 93 books. We perform the following data processing steps to generate aligned data and extract the audio attributes.

- *Transcription* — We first split the audiobook into fragments using silence for segmentation. For each segment, we attempt to generate a transcript using Google Speech Recognition [2]. Note that this transcript is noisy, and we only use it to align with the appropriate chapter.
- *Alignment* — We use the Gentle forced aligner[3] to generate alignments between text and audio, including timestamp and phoneme information at word-level granularity.
- *Audio Attribute Extraction* — To extract the pitch (fundamental frequency in Hz) and the volume (intensity in dB), we use the Python library Parselmouth[4], with the timestep set to 0.01 second (compatible with the granularity of the alignment output from Gentle). To extract the rate of speech, we first compute the z-scores of the durations of all occurrences for each phoneme. Then we compute the mean of the phoneme duration z-scores for each audio segment (phrase / sentence), and the z-score of the resultant means.
- *Sentence Segmentation* — For each sentence, we generate a constituency parse tree using the Berkeley Neural Parser

[2]https://pypi.org/project/SpeechRecognition/
[3]https://github.com/lowerquality/gentle
[4]https://parselmouth.readthedocs.io/en/stable/



(a) Pitch



(b) Volume

Fig. 3: Pitch and volume breakdown for the primary and secondary characters in each book, colored by gender pairs.

[17] with SpaCy [11]. We then split the sentence into phrases, using the constituent sentence (S) components and punctuation.

IV. Audiobook Reader Behavioral Analysis

We compare pitch and volume for the most frequent and the second most frequent characters in each book. Figure 3 shows pitch and volume of character pairs colored by gender. When the two main characters are of opposite gender, female character has a higher pitch and a lower volume compared to male character as most red dots are below the regression line, and most purple dots are above on both plots.

We also compare the pitch and volume for dialogue utterances averaged separately across all the male and female characters in a book. For this comparison, we restrict our analysis to 62 books in which male and female characters each have at least 100 combined word utterances. For 52 out of 62 books, the mean pitch is higher for female character dialogue as compared to male (binomial p=3e-8). For 32 out of 62 books, the mean intensity (volume) is higher for male character dialogue as compared to female (binomial p=0.449).

V. Predicting Prosody Attributes

We now address the task of predicting prosody attributes (pitch, volume, and rate) for each sentence or phrase, given a chapter text as input. We use Mean Squared Error (MSE) between human readers' actual prosody values and our predicted prosody values as accuracy measurement.

A. Text Input Representations

We experiment with three input embeddings: TF-IDF representations, Mean-pooled GloVe embeddings computed from the CommonCrawl pre-trained GloVe model (840B tokens), and MPNet embeddings [19] (`all-mpnet-base-v2`). We compute a single value for each prosody attribute per text segment, and compute the z-scores across all segments in the chapter.

We use a train-test split of 75-25% across books, i.e. we use 69 books (1,392 chapters) in the training set and 24 books (414 chapters) in the test set. All the models are trained on a 2GHz CPU.

B. Joint Prediction

Here we present the test metrics for various models at the phrase level, where we train a single common model to predict all three prosody attributes.

Using only a single phrase embedding as input (with no contextual information about the neighboring phrases), we predict the three attributes. Table I shows the test MSE for the three attributes.

- **LinReg:** We use linear regression as the baseline model, with the default parameters specified in sklearn[5].
- **MLP:** We use the MultiLinear Perceptron with the default parameters specified in sklearn[6], with hidden layer sizes of (5,5), (10,10) and (20,20) respectively.

Embedding	Model	Pitch	Volume	Rate
TF-IDF	LinReg	0.9217	0.8789	0.8439
	MLP (5, 5)	0.9103	0.8661	0.8264
	MLP (10, 10)	0.9080	0.8609	0.8106
	MLP (20, 20)	0.9101	0.8618	0.8067
GloVe	LinReg	0.9197	0.8644	0.9225
	MLP (5, 5)	0.9173	0.8604	0.8905
	MLP (10, 10)	0.9120	0.8488	0.8695
	MLP (20, 20)	0.9014	0.8474	0.8615
MPNet	LinReg	0.8803	0.8048	0.8122
	MLP (5, 5)	1.0001	1.0000	1.0000
	MLP (10, 10)	**0.8667**	0.7817	0.7733
	MLP (20, 20)	0.8668	**0.7798**	**0.7666**

TABLE I: Test MSE for non-contextual joint prediction of prosody attributes

C. Sequential Prediction

We use an LSTM model with input sequence length 2 or 3, with the following architecture: a Bidirectional LSTM layer (size 40), tanh-activation Dense layer (size 20), linear-activation (size 3). We use Mean Squared Error (MSE) as the loss function. We use a validation split of 15%, train for 30 epochs, and select the model with the least validation loss.

[5]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[6]https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

Embedding	Model	Pitch	Volume	Rate
GloVe	LSTM(len 2)	0.8897	0.8260	0.8358
MPNet	LSTM(len 2)	0.8387	0.7540	0.7488
	LSTM(len 3)	**0.8362**	**0.7518**	**0.7449**

TABLE II: Test MSE for contextual joint prediction of prosody attributes with LSTM

Table II shows the test MSE for the three attributes. The LSTM model trained on MPNet phrase embeddings, with sequence length 3 shows the best performance. We use this model further downstream for text-to-speech audio generation and human evaluation.

D. Prediction on Character Dialogue

We observe that human readers read character dialogues more expressively than descriptive text.

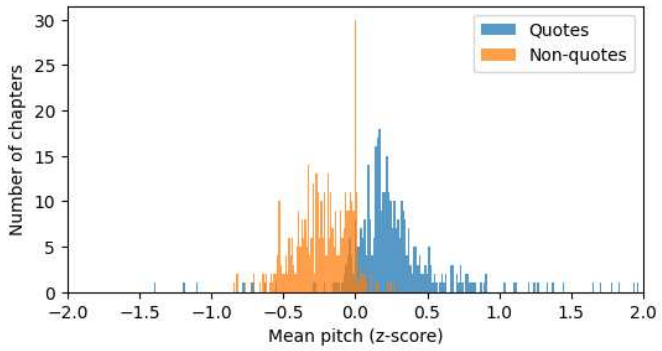


Fig. 4: Mean ground-truth pitch z-score for quote and non-quote phrases from chapters in the test set

Figure 4 shows the distribution of the mean pitch z-scores for phrases that contain quotes or are a part of quotes, and phrases that do not contain quotes. Readers tend to use a higher pitch when reading dialogue and a lower pitch when reading descriptive text, as shown by the right and left shifts from 0 in the distribution. Figure 5 shows the same distributions for our

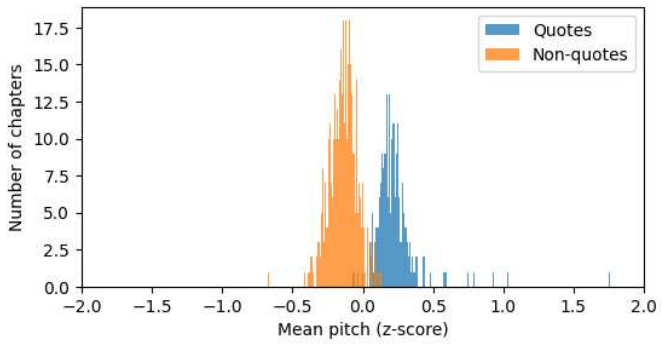


Fig. 5: Mean predicted pitch z-score for quote and non-quote phrases from chapters in the test set

model predictions. These follow a similar pattern, indicating that the model captures information about dialogue, even though we do not explicitly use this information during training.

We further examine this effect by looking at the prosody attribute prediction on dialogues only. Table III shows the MSE on a subset of the test data containing only dialogues. Predicting pitch and volume attributes on dialogues has a lower error compared to all input.

Embedding	Model	Pitch	Volume	Rate
GloVe	LSTM(len 2)	0.8568	0.7683	**0.7773**
MPNet	LSTM(len 2)	0.8073	0.6881	0.8291
	LSTM(len 3)	**0.8018**	**0.6872**	0.8271

TABLE III: Test MSE for contextual joint prediction of prosody attributes on dialogues

E. Character Embeddings

To include more subtle information about the nature of individual characters (e.g. hero vs. villain, old vs. young) that may inform actors' readings, we appended 908-dimensional character embeddings [12] to the input text representation and applied PCA for dimension reduction. Table IV shows the MSE on plain audio generated without SSML attributes (baseline) and enhanced audio generated with SSML attributes predicted from text and character embeddings. Appending character embeddings to the input results in improvements over the baseline, but not over our best models.

Embedding (Size)	Model	Pitch	Volume	Rate
Baseline	N/A	1.3543	1.9550	2.3218
GloVe + CE (200)	LSTM(len 3)	1.0573	0.9559	1.2786
GloVe + CE (100)	LSTM(len 3)	0.9181	0.8021	1.3043
GloVe + CE (50)	LSTM(len 3)	0.9163	0.8169	1.4364
MPNet + CE (200)	LSTM(len 3)	0.9384	0.8368	1.2711
MPNet + CE (100)	LSTM(len 3)	0.8736	**0.7631**	**1.0610**
MPNet + CE (50)	LSTM(len 3)	**0.8734**	0.7709	1.0860

TABLE IV: Test MSE for joint prediction of prosody attributes on dialogues using character embeddings (CE) different PCA dimensions.

VI. Generating and Evaluating Text-to-Speech Audiobooks

We use the best performing model (LSTM with sequence length 3, and MPNet embeddings as input) to generate predictions for each book in the test set. We use a sliding window for inference, and compute the final prediction for each text input as the mean of its prediction in all windows. To convert the predicted z-scores into SSML attributes, we compute the mean and standard deviation for the pitch and volume in each human-read audio file in the test set, and use these values to convert the prosody attributes into absolute values (Hz for pitch and dB for volume). Using the mean pitch/volume as reference, we convert these into relative values as required by the SSML specification (as described in Section II-B). For rate of speech, we use a fixed value for mean and

standard deviation (100 and 50 respectively) and perform the same computation.

We perform a human evaluation using Amazon Mechanical Turk. For each of the 24 books in the test set, we identify the phrases with the highest absolute values of predictions across the three prosody attributes. We then generate audio for this phrase text along with one context phrase before and after, using the Google Cloud Text-to-Speech API [7] with the voice `en-US-Standard-A`. We produce two different audio variants for each text:

- **Plain:** We enclose the text in a `<speak>` tag, without additional prosody information.
- **SSML-enhanced:** We enclose each phrase in a `<prosody>` tag with the three predicted attributes specified: see example in Figure 2.

For each pair of audios, we ask five annotators to rate which audio sounds better and more natural with intonations similar to human readers. Figure 6 shows the results of the evaluation. For 12 out of 24 test sample pairs, a majority of the annotators preferred the SSML-enhanced version.

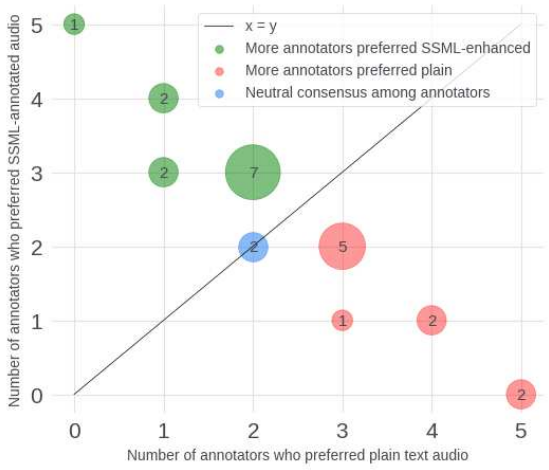


Fig. 6: Distribution of books over annotator preference for SSML or Plain text.

VII. Conclusion

We have developed an aligned text-to-audio book corpus, and deployed it to study both reader behavior and improve prosody models for TTS generation of audiobooks. We have shown that actors customize voices to represent specific characters, and adopt more expressive voices than those generated by conventional TTS systems, motivating the development of full-book analysis methods to understand/predict a reader's goals in narrating a text.

References

[1] D. Bamman, T. Underwood, and N. Smith, "A Bayesian mixed effects model of literary character," 2014. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

[2] J. Brooke, A. Hammond, and G. Hirst, "GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus," 2015. In Proceedings of the Fourth Workshop on Computational Linguistics for Literature, pages 42–47, Denver, Colorado, USA. Association for Computational Linguistics.

[3] Pethe, Charuta, Allen Kim, Rajesh Prabhakar, Tanzir Pial, and Steven Skiena. "STONYBOOK: A System and Resource for Large-Scale Analysis of Novels." arXiv preprint arXiv:2311.03614 (2023).

[4] C. Pethe, A. Kim, and S. Skiena, "Chapter Captor: Text Segmentation in Novels," 2020. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8373–8383, Online. Association for Computational Linguistics.

[5] A. Kim, C. Pethe, and S. Skiena, "What time is it? temporal analysis of novels," 2020. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9076–9086, Online. Association for Computational Linguistics.

[6] D. Wilmot and F. Keller, "Memory and knowledge augmented language models for inferring salience in long-form stories," 2021. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 851–865, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[7] P. Papalampidi, F. Keller, L. Frermann, and M. Lapata, "Screenplay summarization using latent narrative structure," 2020. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1920–1933, Online. Association for Computational Linguistics.

[8] T. Otake, S. Yokoi, N. Inoue, R. Takahashi, T. Kuribayashi, and K. Inui, "Modeling event salience in narratives via barthes' cardinal functions," 2020. In Proceedings of the 28th International Conference on Computational Linguistics, pages 1784–1794, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[9] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," 2014. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

[10] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. Manning, "Stanza: A python natural language processing toolkit for many human languages," 2020. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101–108, Online. Association for Computational Linguistics.

[11] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.

[12] N. Inoue, C. Pethe, A. Kim, and S. Skiena, "Learning and evaluating character representations in novels," 2022. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1008–1019, Dublin, Ireland. Association for Computational Linguistics.

[13] A. Xanthos, I. Pante, Y. Rochat, and M. Grandjean, "Visualising the dynamics of character networks," 2016. In Digital Humanities, pages 417–419.

[14] M. Azab, N. Kojima, J. Deng, and R. Mihalcea, "Representing movie characters in dialogues," 2019. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 99–109, Hong Kong, China. Association for Computational Linguistics.

[15] F. Brahman, M. Huang, O. Tafjord, C. Zhao, M. Sachan, and S. Chaturvedi, ""Let your characters tell their story": A dataset for character-centric narrative understanding," 2021. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[16] P. Taylor and A. Isard, "Ssml: A speech synthesis markup language," 1997. Speech communication, 21(1-2):123–133.

[17] N. Kitaev and D. Klein, "Constituency parsing with a self-attentive encoder," 2018. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

[18] Z. Fu, K. Wang, W. Xin, L. Zhou, S. Chen, Y. Ge, D. Janies, D. Zhang, "Detecting Misinformation in Multimedia Content through Cross-Modal Entity Consistency: A Dual Learning Approach" (2024). PACIS 2024 Proceedings.

[19] K. Song, X. Tan, T. Qin, J. Lu, and T. Liu, " Mpnet: Masked and permuted pre-training for language understanding," 2020. Advances in Neural Information Processing Systems, 33:16857–16867.

[7] https://cloud.google.com/text-to-speech