Evaluating Question Prosody in Text-to-Speech Software

Alexandra C. Salem[1], *Sarah Ita Levitan*[1,2,3]

[1]The Graduate Center at the City University of New York, New York, NY, USA
[2]Hunter College, New York, NY, USA
[3]Bar Ilan University, Ramat-Gan, Israel

asalem1@gradcenter.cuny.edu, sarah.levitan@hunter.cuny.edu

Abstract

Text-to-speech (TTS) software has improved dramatically in recent years after the success of several machine learning innovations resulting in the current AI surge. In spite of its popularity, TTS still produces some infelicitous utterances in simple sentence structures, such as questions. In Mainstream American English, Wh- questions typically have falling pitch, while Yes-No questions have rising pitch. We evaluate how well TTS produces these expected patterns on questions that occur in a short story, and compare with a human orator audiobook. We test two modern TTS systems, Kokoro and Fastspeech 2. We use automated methods to label pitch rise or fall at the end of each question. Kokoro, the better performing model, produced rising pitch for the Yes-No questions 66.7% of the time, compared with 77.8% by the human orator, and rising pitch for the Wh-questions 61.9% of the time, compared with 31.6%. These results demonstrate some mismatch between TTS and the human orator, and indicate that further work is needed to synthesize speech with naturalistic prosody.

Index Terms: text-to-speech, speech synthesis, question prosody

1. Introduction

The task of Text-to-Speech (TTS)—also known as speech synthesis—is the process of taking text as input and producing naturalistic sounding speech using automated methods. TTS software has improved dramatically in recent years after the success of recent deep learning innovations such as the transformer architecture [1] (the basis of Large Language Models). These modern TTS systems are typically evaluated with human qualitative evaluations, where participants listen to samples and provide Mean Opinion Scores (MOS) on the quality of the speech. Modern systems such as Tacotron2 achieve MOS as high as 4.5 on a 5 point scale [2]. However, less attention is paid to the more fine-grained characteristics of the speech, such as how felicitous (i.e., natural) the prosody is. In this work, we specifically evaluate aspects of prosody in speech synthesized with modern TTS software.

Some recent papers have introduced TTS systems that allow global control over certain prosodic features like duration or spectral tilt [3, 4]. Most similar to our work, others have developed automated pitch and phrase accent prediction [5], and evaluated training a TTS system with the prosody-labeled text [6]. However, this work used a Merlin-based TTS system [7], which is now outdated in comparison to recent TTS models. To our knowledge, no recent studies have evaluated fine-grained prosody of synthesized speech from modern end-to-end TTS systems.

Thus, our goal was to evaluate how well modern TTS systems produce appropriate and natural prosody in comparison to a human orator. We focus on a simple type of sentence with a known prosodic pattern: *questions*. In Mainstream American English (MAE), there are two main question-types with distinct prosodic characteristics: Wh- questions (i.e., questions starting with *who*, *what*, *when*, *where*, and *how*), and Yes-No questions (i.e., questions that elicit a response of *yes* or *no*).

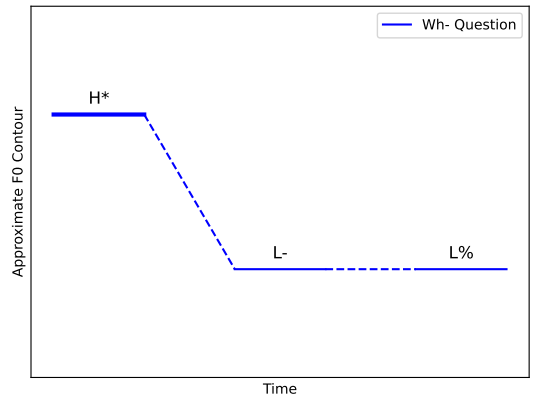
It is generally reported that Wh- questions have declarative intonation, including falling pitch at the end of the question. Most commonly, the well-known Wh- question pitch contour is a "high-fall" contour which has a ToBI (Tones and Break Index) transcription of H*L-L% [8]. That is, the final nuclear contour has a high tone pitch accent, and ends with a low phrase accent and low boundary tone. This contour is represented in Figure 1a. In a 2010 report from Hedberg et al. [9], they report that this exact contour actually occurs for only 49% of Wh- contours in their sample, followed by 25% of a similar contour, L+H*L-L%. Both of these contours, however, share a falling pitch, and thus 74% of Wh- questions contained falling pitch.

In contrast, Yes-No questions are commonly said to have an interrogative contour, characterized by rising pitch at the end of the question. The commonly cited Yes-No contour is the "low-rise" contour with the ToBI transcription L*H-H%, consisting of a final nuclear contour with a low tone pitch accent, a high phrase accent, and a high boundary tone. This contour is represented in Figure 1b. In a 2017 study by Hedberg et al. [10], the authors found that Yes-No questions had this contour, or a highly similar one, 79% of the time.

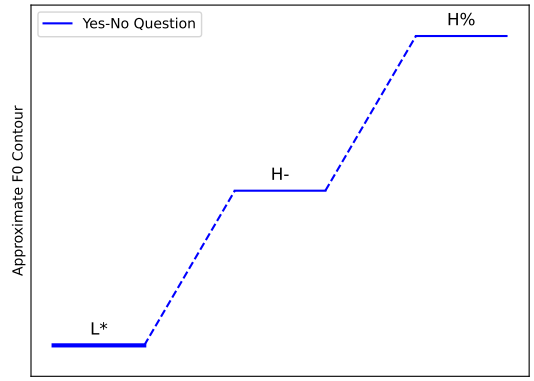
Since these two question types have fairly distinct prosodic patterns, we decided to evaluate prosody in TTS software by investigating whether questions in the synthesized speech followed these same patterns. That is, we wanted to explore whether most synthesized Wh- questions would have falling pitch and most synthesized Yes-No questions have rising pitch.

Audiobooks are an area with significant need for TTS, especially for accessibility. Audiobooks improve accessibility of books not only for people with vision impairment, but also people with ADHD, aphasia, and anyone who struggles with reading comprehension [11, 12]. However, people typically prefer human voices over TTS because they are much more natural and express the emotions that are necessary to capture the full character of a story [13]. Therefore, TTS for audiobooks is an area where improvement in prosody can be impactful.

Thus, in this work, we evaluate the performance of TTS on a constrained set of sentences (Wh- and Yes-No questions), drawn from an audiobook. We choose an audiobook for a short story read by a human orator, so as to limit the scope and prevent any impact of longer-term context in a long text. We generate a parallel set of questions using two TTS systems to read aloud the same audiobook, and evaluate prosody with an automated approach for classifying prosodic pitch patterns. Our research questions are as follows: 1) Can we automatically classify pitch



(a) *Wh- Question*



(b) *Yes-No Question*

Figure 1: *Schematics of F0 contours for two question types*

rise and fall in a way that aligns with human annotators?, 2) Using our automated approach, do we see similar pitch patterns to those reported in the literature for the human orator's question prosody?, and 3) Does TTS software replicate these same pitch patterns for questions in synthesized speech of the same text?

2. Methodology

2.1. Data

We use a short story and corresponding audiobook as our source of question sentences for this work. For our short story we use the fairytale *The Fir Tree* by Hans Christian Andersen from the LibriSpeech corpus [14], a collection of text-audio pairs read my amateur orators. It was read aloud by Taylor Burton-Edward, a speaker of Mainstream American English. We chose this text as our source of questions for two reasons: 1) the story occurs in the test set from LibriSpeech, increasing the chances for it not to be seen during training of TTS models, and 2) due to being a standalone short story, there are no concerns about longer-term context affecting the prosody of the orator.

We obtained the text for the story from a version of LibriSpeech adapted for training TTS software, LibriTTS [15]. Importantly, LibriTTS restores the punctuation to the text from LibriSpeech, allowing us to automatically identify the questions that occur in the text. The text in LibriTTS is also split at the sentence-level as opposed to paragraph-level, providing start and end times for each sentence. However, this process was done automatically and made several mistakes, and thus the first author went through and re-segmented each question using Praat software [16]. Specifically, each question was split not at the full sentence-level, but at the actual question-level. For instance, in the sentence "Where are they going to? asked the Fir.", the audio was segmented at "Where are they going to?".

There were 33 total questions in the story. We labeled each question as a Wh- question or a Yes-No question. Two sentences did not quite fit in either category, leaving 31 questions. There are 22 Wh- questions and 9 Yes-No questions.

2.2. Text-to-Speech Models

We tested two popular open-source TTS systems: Kokoro [17] and Fastspeech 2 [18]. Both models take text as input and generate a Mel spectrogram which is then transformed to audio using a vocoder. Each of these models generate speech efficiently and are available on HuggingFace.[1]

Kokoro is a lightweight open-weight TTS model. It uses StyleTTS 2 [19], an end-to-end system consisting of eight modules (such as a text encoder, prosody predictor, and pitch extractor) to transform input text into Mel spectrograms, followed by iSTFTNet [20], a vocoder to transform to audio. The exact dataset used for model training is not provided, but the authors use only non-copyrighted audio data produced by humans. Kokoro is one of the most frequently used TTS systems available on HuggingFace.[2]

Fastspeech 2 is a feed-forward transformer-based model incorporating a variance adapter for pitch, duration, and energy determination. The implementation we used incorporates HiFi-GAN for the vocoder to transform the generated Mel spectrograms to audio [21]. It was trained on LJSpeech, a corpus of non-fiction read by a single speaker [22].

2.3. Evaluation

Our goal was to identify similarities and differences in prosody for questions from human orators and TTS systems. In an ideal world, we would evaluate the speech by transcribing the prosody in ToBI. However, ToBI transcription is incredibly labor intensive, and further it is not known how well it can be annotated for synthesized speech. Instead, we used automated tools to approximate some characteristics of prosody. We evaluated two specific characteristics of prosody for the questions: 1) proportion of questions with rising F0 at the end of the question, and 2) average F0 range across questions. We calculated each of these metrics separately for Wh- and Yes-No questions, for both the human orator and two TTS systems. We also validated our method for classifying rising F0, and conducted a qualitative evaluation of question prosody.

2.3.1. Automatic Evaluation of Question Prosody

To evaluate rising versus falling F0, we first obtained F0 contours for each question audio clip using the pYIN algorithm, implemented in the Librosa Python package [23, 24]. Then, for the latter half of each question, we used least squares polynomial fit to determine a Line-of-Best-Fit (LoBF) of the F0 contour. We chose to use only the second half of the audio since it is possible for there to be additional high and low tones throughout a

[1] https://huggingface.co/

[2] As searched on https://huggingface.co/ 2025-12-04.

question, and thus looking at the latter half made it more likely to capture the appropriate final pattern. During LoBF determination, any unvoiced frames with undefined F0 were removed from the calculation. If the resulting slope for the LoBF was negative, we classified the pitch for that question as rising. Otherwise, we classified the pitch for that question as falling. We then aggregated these results across audio source and question type to obtain the proportion of questions with rising F0.

To calculate average pitch range, we obtained the maximum and minimum F0 for each question, calculated their difference, and averaged across question types.

2.3.2. *Validation of Automatic Prosody Identification*

Capturing whether a LoBF for the F0 has positive or negative slope is an imperfect method to capture the true measure we are after: whether the pitch rises or falls at the end of the question, reflected in particular ToBI transcriptions of the nuclear contour. Not only is pitch a subjective concept distinct from F0 itself, but also F0 can be undefined when sounds are unvoiced, or the slope of F0 could be rising overall, but nonetheless have lower pitch at the very end. Thus, to verify the validity of this automated labeling approach, we compared its results with a selection of questions for which we hand-annotated ToBI transcriptions for the nuclear contour. Two trained ToBI transcribers annotated 20 questions (14 Wh-, 6 Yes-No) from the human orator. The following transcriptions were marked as rising: L*H-H%, L*H-L%. These transcriptions were marked as falling: H*L-L%, L+H*L-L%, L+H*L-H%, !H*H-L%. The resulting human labels were compared with the LoBF labels using Cohen's kappa [25].

2.3.3. *Qualitative Human Evaluation of Question Prosody*

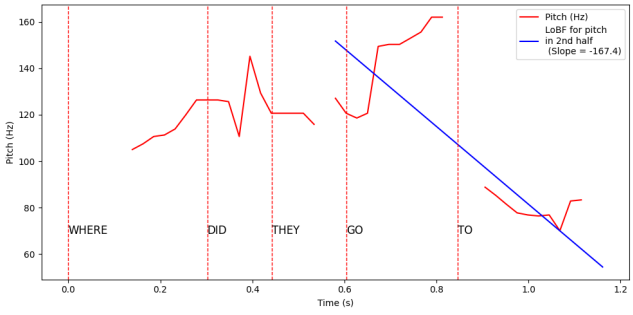
Finally, we conducted a preliminary qualitative evaluation of the naturalness of the prosody for questions in the synthesized speech. We randomly selected 11 questions from the story. We recruited three human reviewers who were native speakers of MAE, without specific expertise in prosody. We split the reviews such that each question from each source (human, Kokoro, Fastspeech 2) was rated by two reviewers. We asked each reviewer to rate whether each question was *natural* using a Likert Scale from Strongly Disagree to Strongly Agree, which we then coded as 1-5. Before starting the survey, the reviewers were told to pay particular attention to whether the prosody sounded natural, and they were given an example of a both a natural and infelicitous Wh- question and Yes-No question.

3. Results

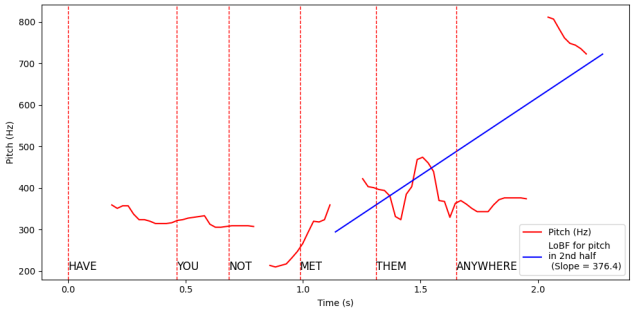
3.1. Automatic Evaluation of Question Prosody

Examples of F0 contours and their corresponding LoBF for a Wh- question and a Yes-No question are shown in Figures 2a and 2b. In each of these examples, our calculated LoBF slope matched the pattern shown in the ToBI transcriptions.

Our automated pitch metric characteristics – proportion of questions with rising F0, and average range of F0 – are shown in Table 1. We found that for the human orator, 31.6% of Wh-questions had rising F0, and 77.8% of Yes-No questions had rising F0, which aligns with expected patterns. The speech synthesized with Fastspeech 2 saw around the same proportion of rising F0 for both Wh- and Yes-No questions – 42.9% and 44.4%, respectively. Kokoro actually used rising F0 more often than falling F0 (61.9% of the time) for Wh- questions, and a slightly higher percentage for Yes-No questions (66.7% of the time). Looking at average F0 range, the human orator had the highest range with 183.3 for Wh- questions and 396.8 for Yes-No questions. Kokoro had ranges of 131.3 for Wh- and 120.2 for Yes-No. Fastspeech 2 had the lowest ranges of 87.3 for Wh-questions and 99.8 for Yes-No questions.



(a) *Wh- Question with ToBI Transcription H*L-L%*



(b) *Yes-No Question with ToBI Transcription L*H-H%*

Figure 2: *Actual F0 contours and corresponding Lines-of-Best-Fit (LoBF) of two example questions from the human orator*

Table 1: *Prosody characteristics for human orator and TTS*

Source	Question Type	Prop.(%) Rising F0	Avg. F0 Range
Human orator	Wh-	31.6%	183.3
TTS – Fastspeech 2	Wh-	42.9%	87.3
TTS – Kokoro	Wh-	61.9%	131.3
Human orator	Yes-No	77.8%	396.8
TTS – Fastspeech 2	Yes-No	44.4%	99.8
TTS – Kokoro	Yes-No	66.7%	120.2

3.2. Validation of Automatic Prosody Identification

Our evaluation of agreement between LoBF pitch slope prediction and human transcriber pitch rising determination is shown in Table 2. This analysis includes the 16 out of 20 transcribed questions which the two human transcribers agreed upon. 11 out of 16 were Wh- questions. For these questions, the human transcribers and the LoBF analysis agreed on pitch rising status for nine out of 11 questions or 81.8% of the time. Two questions were labeled without rising pitch by the human transcribers, which the LoBF analysis labeled as having rising pitch. Only one question was labeled with rising pitch by the human transcribers, but labeled as not having rising pitch using LoBF slope. For Yes-No questions, we were left with a small sample of just five questions. However, four out of five of these (80%)

saw agreement between the human transcriber and LoBF slope analysis. Overall, the agreement between human transcribers and LoBF F0 slope for pitch rising label was 13 out of 16 or 81.3%. The Cohen's kappa value was 0.589, which is typically considered "moderate" agreement [26]. Though not perfect, we took this level of agreement to indicate that the F0 slope captured the appropriate nuclear contour characteristic sufficiently well enough for our purposes.

		LoBF	
	Human Transcriber	Pitch rising	
Question Type	Pitch rising	Yes	No
Wh-	**Yes**	1	0
	No	2	8
Yes-No	**Yes**	3	1
	No	0	1

Table 2: *Confusion matrices between human transcription of pitch rising and LoBF determination of pitch rising, for Wh- and Yes-No questions.*

3.3. Qualitative Human Evaluation of Question Prosody

Finally, results from our qualitative survey on question naturalness are shown in Table 3. As expected, the reviewers rated the human orator's questions as the most natural, with an average of 4.227 on a scale of 1-5. Fastspeech 2 had the lowest rating, and a large variability, with an average of 1.909 and a standard deviation of 0.861. Kokoro achieved a fairly high naturalness rating of 3.546, putting the reviewer's ratings of whether it was natural between "Neutral" and "Agree".

Source	Naturalness	
	Avg	SD
Human orator	4.227	0.607
TTS – Fastspeech 2	1.909	0.861
TTS – Kokoro	3.546	0.789

Table 3: *Qualitative evaluation of question naturalness (range 1-5) across reviewers*

4. Discussion

It has generally been understood that in Mainstream American English Wh- questions have falling pitch—declarative intonation—while Yes-No questions have rising pitch. These patterns were observed in our results, where the human orator used rising F0 31.6% of the time for Wh- questions and 77.8% of the time for Yes-No questions.

Kokoro TTS software did produce more Yes-No questions with rising F0 than falling F0 (66.7%), similar to (though lower than) the pattern seen for the human orator (77.8%). However, Kokoro also produced rising pitch for 61.9% of Wh- questions, which does not align with the human orator. We found that non-expert human listeners showed weak agreement with the statement that Kokoro produced natural-sounding questions. Additionally, Kokoro speech was also fairly variable in its F0 range, possibly indicating more expressive speech. Overall, this system demonstrated learning some patterns of human question prosody, but there was clearly room for improvement.

Fastspeech 2 produced rising F0 42.9% of the time for Wh- questions and 44.4% of the time for Yes-No questions. Thus, this system did not really distinguish between question types at all. Fastspeech 2 also produced a much more limited F0 range than either the human orator or Kokoro for both question types, indicating a lack of variability or expressiveness of the speech. Moreover, the reported naturalness from the qualitative survey showed low naturalness of 1.9—just below choosing "Disagree" that the question is natural. Across each of these metrics, it is clear that despite the overall quality of the produced speech, this system did not produce the expected prosody patterns for questions in a short story.

For this work, we only considered open-source models due to concerns about reproducibility. However, it is possible that proprietary systems such as Elevenlabs or Naturalreader, which are highly popular, could produce questions with prosodic patterns that are more similar to humans. Thus, an area of future work is to explore these commercial systems. Additionally, we only tested two open-source models. We chose these models due to their efficiency and popularity, but there are numerous TTS systems to choose from, some of which could have had higher performance. Notably, Fastspeech 2 was trained on non-fiction, and thus may not be the ideal choice for reading aloud a story that aims to be entertaining. In future work, additional open-source TTS systems, including ones that are specifically designed for reading aloud fiction, should be explored.

We chose the short story *The Fir Tree* due to its limited size, American orator, and existence in the test set of LibriSpeech. However, there are some idiosyncrasies of this story. While the reader speaks MAE generally, there are points in the story where he uses playful voices for characters that demonstrate non-MAE characteristics. Additionally, he often uses a falsetto voice for characters, which leads to more difficulty with ToBI transcription. In future work, additional stories that do not have these idiosyncrasies should be explored.

In this work, we avoided using terms such as "accuracy" or "correctness" of the question prosody in TTS, since not only is our main metric for capturing rising pitch imperfect (80% agreement with human annotators), but also humans themselves do not always produce questions with the expected prosody. Yet, the differences between the TTS systems in comparison to the human orator are pronounced. These comparisons indicate that in spite of overall high MOS of TTS systems, there is still quite a bit of room for improvement. The TTS systems seem to have learned something about prosody, due to at least Kokoro's relatively high qualitative naturalness rating. But, it is clear that these systems have not learned to consistently replicate Wh- and Yes-No question patterns. These models learn prosodic patterns in a data-driven manner and there is no direct prosody-modeling in their training. While this data-driven approach clearly offers some success, the issues with fine-grained prosodic characteristics indicate it may be helpful to incorporate prosodic information into the training.

5. Conclusions

TTS systems have seen remarkable progress in recent years, but our job is not done. While MOS scores are helpful for capturing the overall quality of synthesized speech, it is important to explore fine-grained performance as well. This work offers a preliminary analysis in that direction by focusing on questions, which have known prosodic patterns. Overall, we find that one of our tested TTS systems—Kokoro—demonstrates some understanding of prosodic patterns in questions for MAE. However, more work is needed to achieve TTS performance that is truly felicitous.

6. Acknowledgments

No funding disclosure. We thank Jason Bishop for his transcriptions of our monstrous questions as well as Kei Kereau, Gabe Levine, and a third unnamed participant for taking our qualitative survey. We also thank Kyle Gorman for his contributions to the initial ideas for this paper.

7. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB: IEEE, Apr. 2018, pp. 4779–4783. [Online]. Available: https://ieeexplore.ieee.org/document/8461368/

[3] S. Latif, I. Kim, I. Calapodescu, and L. Besacier, "Controlling prosody in end-to-end TTS: A case study on contrastive focus generation," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, A. Bisazza and O. Abend, Eds. Online: Association for Computational Linguistics, Nov. 2021, pp. 544–551. [Online]. Available: https://aclanthology.org/2021.conll-1.42/

[4] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 4432–4436. [Online]. Available: https://www.isca-archive.org/interspeech_2020/raitio20_interspeech.html

[5] R. Sloan, S. S. Akhtar, B. Li, R. Shrivastava, A. Gravano, and J. Hirschberg, "Prosody prediction from syntactic, lexical, and word embedding features," in *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 269–274.

[6] R. Sloan, A. Adigwe, S. Mohandoss, and J. Hirschberg, "Incorporating prosodic events in text-to-speech synthesis," in *Speech Prosody 2022*, 2022, pp. 287–291.

[7] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 202–207.

[8] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The Original ToBi System and the Evolution of the ToBi Framework," in *Prosodic Typology*, 1st ed., S.-A. Jun, Ed. Oxford University PressOxford, Jan. 2005, pp. 9–54. [Online]. Available: https://academic.oup.com/book/4251/chapter/146095715

[9] N. Hedberg, J. M. Sosa, E. Görgülü, and M. Mameni, "The prosody and meaning of wh-questions in American English," in *Speech Prosody 2010*. ISCA, May 2010, pp. paper 045–0. [Online]. Available: https://www.isca-archive.org/speechprosody_2010/hedberg10_speechprosody.html

[10] N. Hedberg, J. M. Sosa, and E. Görgülü, "The meaning of intonation in yes-no questions in American English: A corpus study," *Corpus Linguistics and Linguistic Theory*, vol. 13, no. 2, pp. 321–368, Sep. 2017. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/cllt-2014-0020/html

[11] K. J. Esteves and E. Whitten, "Assisted reading with digital audiobooks for students with reading disabilities," *Scholarship and Professional Work – Education*, no. 75, 2011. [Online]. Available: https://digitalcommons.butler.edu/coe_papers/75

[12] K. K.-P. Sarah E. Wallace and K. Hux, "How text-to-speech aids reading for people with aphasia," *ASHA Leader*, 2023. [Online]. Available: https://leader.pubs.asha.org/do/10.1044/leader.FTR1b.28032023.slp-text-to-speech.52/full/

[13] 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*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–13. [Online]. Available: \url{https://doi.org/10.1145/3313831.3376789}

[14] 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, pp. 5206–5210.

[15] 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 *Interspeech 2019*, 2019, pp. 1526–1530.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2025. [Online]. Available: https://praat.org

[17] Hexgrad, "Kokoro-82m (revision d8b4fc7)," 2025. [Online]. Available: https://huggingface.co/hexgrad/Kokoro-82M

[18] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," 2020. [Online]. Available: https://arxiv.org/abs/2006.04558

[19] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[20] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform," in *ICASSP*, 2022.

[21] J. Kong, J. Kim, and J. Bae, "Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[22] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[23] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:3287290

[24] B. McFee, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, Emily Halvachs, Carl Thomé, Fabian Robert-Stöter, Rachel Bittner, Ziyao Wei, Adam Weiss, Eric Battenberg, Keunwoo Choi, Ryuichi Yamamoto, CJ Carr, Alex Metsai, Stefan Sullivan, Pius Friesch, Asmitha Krishnakumar, Shunsuke Hidaka, Steve Kowalik, Fabian Keller, Dan Mazur, Alexandre Chabot-Leclerc, Curtis Hawthorne, Chandrashekhar Ramaprasad, Myungchul Keum, Juanita Gomez, Will Monroe, Viktor Andreevitch Morozov, Kian Eliasi, nullmightybofo, Paul Biberstein, N. Dorukhan Sergin, Romain Hennequin, Rimvydas Naktinis, beantowel, Taewoon Kim, Jon Petter Åsen, Joon Lim, Alex Malins, Darío Hereñú, Stef van der Struijk, Lorenz Nickel, Jackie Wu, Zhen Wang, Tim Gates, Matt Vollrath, Andy Sarroff, Xiao-Ming, Alastair Porter, Seth Kranzler, Voodoohop, Mattia Di Gangi, Helmi Jinoz, Connor Guerrero, Abduttayyeb Mazhar, toddrme2178, Zvi Baratz, Anton Kostin, Xinlu Zhuang, Cash TingHin Lo, Pavel Campr, Eric Semeniuc, Monsij Biswal, Shayenne Moura, Paul Brossier, Hojin Lee, Waldir Pimenta, Shin Hyun, Iliya S, Eugene Rabinovich, Geo Lei, Jize Guo, Phillip S. M. Skelton, Matt Pitkin, Anmol Mishra, Slava Chaunin, BenedictSt, Scott VanRavenswaay, and David Südholt, "librosa/librosa: 0.11.0," Mar. 2025. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.591533

[25] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104

[26] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.