



RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis

Rohola Zandie^{1,2}, Mohammad H. Mahoor^{1,2}, Julia Madsen², and Eshrat S. Emamian²

¹Department of Electrical and Computer Engineering, University of Denver

²DreamFace Technologies, LLC

rohola.zandie@du.edu, mmahoor@du.edu, {jmadson, eemamian}@dreamfacetech.com

Abstract

This paper introduces *RyanSpeech*, a new speech corpus for research on automated text-to-speech (TTS) systems. Publicly available TTS corpora are often noisy, recorded with multiple speakers, or lack quality male speech data. In order to meet the need for a high quality, publicly available male speech corpus within the field of speech recognition, we have designed and created *RyanSpeech* which contains textual materials from real-world conversational settings. These materials contain over 10 hours of a professional male voice actor's speech recorded at 44.1 kHz. This corpus's design and pipeline make *RyanSpeech* ideal for developing TTS systems in real-world applications. To provide a baseline for future research, protocols, and benchmarks, we trained 4 state-of-the-art speech models and a vocoder on *RyanSpeech*. The results show 3.36 in mean opinion scores (MOS) in our best model. We have made both the corpus and trained models for public use.

Index Terms: text to speech, speech corpus, speech recognition

1. Introduction

The advent of end-to-end deep neural networks (DNN) in the fields of automatic speech recognition (ASR) and text-to-speech (TTS) has shifted the research paradigm in these fields. Traditional methods are complex and time-consuming, requiring predefined linguistic features which are typically language specific. However, new TTS architectures are composed of two parts: first, they generate a Mel-spectrogram autoregressively or non-autoregressively from the input text's phonemes, and then output audio from a separately trained vocoder. The quality of the final outputs depends heavily on the quality of speech corpora. Even though researchers [1, 2, 3] in recent years have paid more attention to create larger speech and language corpora, it is evident that more research is necessary to fill the gaps of speech corpora.

This paper addresses shortcomings in recent research and scholarship within the fields of corpus development by introducing *RyanSpeech*, the first publicly available male voice TTS corpus in the conversational setting. Using state-of-the-art deep neural networks, we have found that it is easier to train a model capable of generating natural speech synthesizers than traditional methods [2]. However, this is only possible when we have an available corpus at hand. Most of the publicly available corpora are in the domains of reading or audiobooks, and most of these corpora are not single-speakers, which is unsuitable for training the TTS (especially the vocoder models). To have a high-fidelity TTS, the corpus also needs to be recorded at a high sampling rate without any background noise. LJSpeech is the only female single-speaker corpus with low noise and a non-restrictive license [1]. *RyanSpeech* includes features that make it ideal for a TTS system in real-world applications. These features are:

- *RyanSpeech* is the only corpus in the domain of conversation. Other speech corpora in the domain of conversation are multi-speaker and recorded in high-noise environments that are not appropriate for training TTS systems. In *RyanSpeech*, we extracted the most commonly used conversations with prosodic variations specific to conversations that cannot be replaced by audiobooks or reading settings.
- All the audio files are recorded by a single professional male speaker with studio quality at a sampling rate of 44100 Hz. We double-checked all the recordings and rerecorded sample files that were too slow, too fast, or contained unwanted speech variations. This makes *RyanSpeech* the first male speaker corpus for building high-quality TTS systems.
- We chose the sentences to reflect numerous and diverse real-life conversational situations, including dialogues on movies, sports, music, television, restaurants, and nature as well as common questions and general discussions.
- We open-sourced the corpus including all the original FLAC (Free Lossless Audio Codec) files with transcripts under the CC BY-NC-ND license to help rapid development in TTS research.

This paper is organized as follows. Section 2 summarizes speech corpora, especially those most relevant to our own. Section 3 details the corpus creation pipeline from collecting raw text to the final audio files. Section 4 provides the corpus's overall statistics, including both audio and textual transcripts. We present the details of training different models based on *RyanSpeech* as well as the human evaluation and the results in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

Speech corpora can be categorized into multi- and single-speakers. Multi-speaker corpora are designed to capture the diversity in spoken language with different voices, genders, ages, and accents. These corpora are well-suited to automatic speech recognition systems that require variations in input data [4]. However, creating a high-quality TTS system necessitates single-speaker corpora [2], which is especially important for training the vocoder.

Table 1 demonstrates different corpora with public licenses widely used for research on speech recognition and synthesis. The CMU ARCTIC corpus is a dataset that has been used for years as a baseline for speech recognition tasks [5]. The VCTK corpus contains 110 English speakers with various accents [6]. The main sources for the passages are newspapers selected using a greedy algorithm to increase phonetic diversity. Common Voice is a project by Mozilla which collects the

biggest speech corpus by crowdsourcing from their community [7]. This corpus has the greatest diversity, and is also very noisy. VoxForge is most similar to Common Voice because it is also a community-based project [8]. Unlike Common Voice, VoxForge does not have a verification process in the data collection process. LibriSpeech [3] and LibriTTS [2] derive their text source from the LibriVox project which is based on audiobooks [9]. LibriTTS is the successor of LibriSpeech and has more robust design choices including high-quality sampling rate, the removal of noisy subsets, and split speech at sentence breaks. In the domain of conversational speech corpora, we have CHiME-5 [10] which consists of 20 parties each recorded in different houses. Also, the Linguistic Data Consortium (LDC) has developed CALLHOME American English Speech, which consists of 120 unscripted 30-minute telephone conversations between native English speakers [11]. Both of these corpora are noisy and contain multiple speakers, which renders them not particularly suitable for TTS training.

As we mentioned above, single-speaker corpora are the best candidates for high-quality TTS systems. The BC2013 corpus comes from the blizzard challenge and contains a large amount of speech by a single speaker with good quality [12]. While this corpus is based on Audiobooks, it is not ideal for many applications that require user interaction. M-AILABS was recorded with a male and female voice, but it is too noisy and unsuitable for training the speech synthesizer [13]. The only acceptable quality speech TTS corpus is LJSpeech, which is all recorded by a female voice. *RyanSpeech* is the first high-quality male TTS corpus with a non-restrictive license in the conversational domain that can be used for training both the vocoder and speech model.

3. Corpus creation pipeline

This section describes the data processing which we developed to produce the *RyanSpeech* corpus.

3.1. Data collection

Since *RyanSpeech* is a single-speaker speech corpus designed and developed for conversational systems, we used three different text resources that are most relevant to this task.

1. **Ryan Chatbot dataset:** This is a specifically designed dataset developed for Ryan Robot [14, 15] in different conversational settings. It contains more than 56,000 sentences covering a wide variety of topics, including television, sports, movies, music, science, food, museums, and history. We randomly selected 5778 sample sentences from this dataset, and what follows is a short list of example sentences that appear in our “food” dialogues:
 - “Do you want to see another recipe that you could easily prepare at home?”
 - “I’m coming over to your house, and I’m coming hungry!”
 - “Maybe you just need to have a nice, home-cooked meal.”
2. **Taskmaster-2:** This dataset is also designed for use in goal-oriented dialogue systems [16]. It includes both user and assistant roles in the conversations. Taskmaster-2 consists of 17289 dialogues in the seven domains of restaurants, food ordering, movies, hotels, flights, music, and sports. For *RyanSpeech*, we randomly selected 3000 samples in all categories of Taskmaster-2.
3. **LibriTTS:** To balance the dataset, we also included 2501 random samples from the LibriTTS text corpus. LibriTTS

is a multi-speaker speech dataset with 116500 sentences in their train-clean-360 subset. We randomly selected 2501 sentences from this subset.

3.2. Text pre-processing

After collecting texts from different sources, we underwent sentence segmentation and normalization.

1. **Sentence segmentation:** We used Spacy [17] for sentence segmentation on text from the Ryan Chatbot dataset and Taskmaster-2. It uses a trainable pipeline component for sentence segmentation.
2. **Text normalization:** we detected non-standard words and normalized them to read text manually. In this step, we address numbers (cardinal numbers, signed integers, real numbers, ordinal numbers, roman numerals, fractions, and sequence of digits like phone numbers), Currency, Time, Date, Abbreviations and Street addresses.

3.3. Trim silence

After recording the audio files, we needed to trim the beginning and end silences in each file. For this purpose, we wrote a simple script to trim the audio based on thresholding the sound amplitude. To make it even more accurate, we manually double-checked all the trimmings with Audacity audio software to ensure that we correctly removed the silences.

3.4. Post processing

Even though all the audios had been recorded in a studio with high-quality, we normalized the sound amplitude of recordings to ensure that they all had the same volume level.

4. Statistics of the corpus

RyanSpeech contains 9.84 hours of high-quality audio recorded in a studio by a professional voice actor. The speaker has the US English dialect. The sampling rate of all audio files is 44,100 Hz. We used the audio coding format of FLAC, which is lossless compression.

Figure 1 shows violin plots of the number of characters per sentence in LibriTTS, LJSpeech, LibriSpeech, M-AILABS, and *RyanSpeech*. It can be seen from the figure that *RyanSpeech* has relatively shorter sentences compared to other corpora. The mean sentence length is 58.07 characters with a standard deviation of 26.09, closest to LibriTTS. The distribution of sentence length from *RyanSpeech* to M-AILABS is also significantly different. Our corpus sentence lengths follow a peaked distribution, whereas in M-AILABS, it has a larger standard deviation (std=50.80).

In Figure 2, the Box Plot of audio duration for different corpora are demonstrated. Because our corpus is in the domain of conversation, the audio files are shorter. The main reason for this observation is that the properties of conversational recordings are usually shorter and faster than audiobooks and other reading corpora. The diversity of audio length in terms of standard deviation in *RyanSpeech* is less than other corpora, making it suitable for a stable training corpus for assistant and chat dialogue systems.

5. Experiments

The training pipeline consists of two steps: first, the character level input text is fed into a TTS system that outputs Mel spectrogram frames, then we use a trained vocoder to convert the Mel spectrograms to waveforms.

Table 1: List of the publicly available multi and single speaker speech dataset

Corpus	Domain	Licence	Duration (hours)	Sampling rate (kHz)	Number of Speakers
CMU ARCTIC [5]	Reading	BSD	7	16	7
VCTK [6]	Reading	ODC-By v1.0	44	48	109
Common Voice [7]	Reading	CC-0 1.0	1,118	48	51,072
VoxForge [8]	Reading	GPL	120	16	2966
LibriSpeech [3]	Audiobook	CC-BY 4.0	982	16	2,484
LibriTTS [2]	Audiobook	CC-BY 4.0	586	24	2,456
CHiME-5 [10]	Conversation	Commercial	50	16	48
CALLHOME [11]	Conversation	LDC	60	8	120
BC2013 [12]	Audiobook	Non-commercial	300	44.1	1
M-AILABS [13]	Audiobook	BSD	75	16	2
LJSpeech [1]	Audiobook	CC-0 1.0	25	22.05	1
RyanSpeech	Conversation	CC BY-NC-ND	10	44.1	1

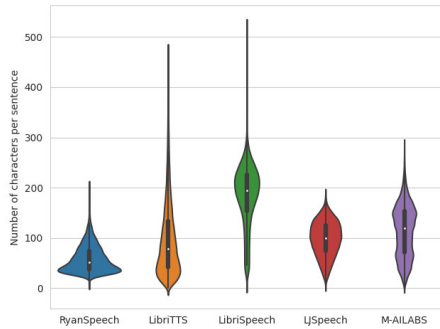


Figure 1: The violin plots of the number of characters per sentence for RyanSpeech, LibriTTS, LibriSpeech, LJSpeech and, M-AILABS. The thick line shows the interquartile range (from 25% to 75%), and the white dot is the median value. The width of the violin plot in any point indicates the frequency.

5.1. Training neural vocoder

The speech synthesizer gives us the time-domain waveform samples from the Mel spectrograms feature representations produced by the TTS system. In the first step, we trained Parallel-WaveGAN [18], which is a fast waveform generation method that uses a non-autoregressive WaveNet model. It is based on Generative Adversarial Network (GAN) and does not need a two-stage teacher-student framework for training. Basically, the model is trained by optimizing multi-resolution Short-time Fourier transform (STFT) on the spectrogram and an adversarial loss function at the same time. For training the feature extractor, the sampling rate is set to 22,050 Hz with the FFT window size of 1024. For the generator, we set the number of residual block layers to 30, the number of dilation cycles to 3, and the number of residual channels to 64. The number of layers for the discriminator network is set to 10. The loss balancing coefficient is also set to 4. The model was trained for 400k steps. Figure 3 shows one example of the generated sample after 400k steps versus the ground truth. It illustrates that the generative network is capable of creating a very similar waveform to the ground truth. The training was done on a NVIDIA TITAN-Xp GPU with 12 GB of memory and a batch size of 6. The training takes 95.74 hours.

5.2. Training text to speech model

We used different architecture for training the TTS system. Tacotron is a RNN-based model that uses CBHG (1-D convolution bank + highway network + bidirectional GRU) module

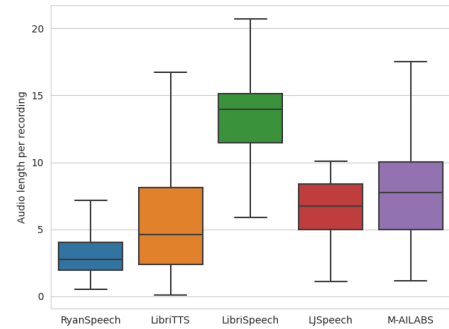


Figure 2: The box plots of audio length for RyanSpeech, LibriTTS, LibriSpeech, LJSpeech and, M-AILABS. The box shows the interquartile range (from 25% to 75%); the middle line shows the median value. The whole range is from the minimum to maximum value. RyanSpeech has the shortest audio length which is a characteristic of conversational dialogue.

Table 2: The comparison between standard deviation (σ), skewness(γ) and kurtosis (κ) of pitch in ground-truth and synthesized audio

Model	σ	γ	κ
Ground Truth	65.64	-0.630	0.256
Tacotron	63.97	-0.677	0.302
FastSpeech	68.61	-0.532	0.165
FastSpeech2	67.71	-0.565	-0.023
Conformer	65.77	-0.703	0.119

[19]. We trained Tacotron for 100k steps. Transformer-based models have been adopted in TTS systems due to their success in modeling long-range dependencies. We trained FastSpeech [20] and FastSpeech2 [21] which are based on transformer. FastSpeech is based on attention mechanism and 1D convolution. These models have a module for length regulation that addresses the problem of mismatch between the phoneme and spectrogram sequence. The main difference between FastSpeech and FastSpeech2 is that the duration predictor for length regulation is trained using a teacher model in FastSpeech, while in FastSpeech2 it is trained end to end, which is faster and more accurate. Finally, we used Conformer [22] which brings the ideas of convolutional networks to the transformer-based model. The main difference between conformer and transformer models is that conformer has two feedforward layers

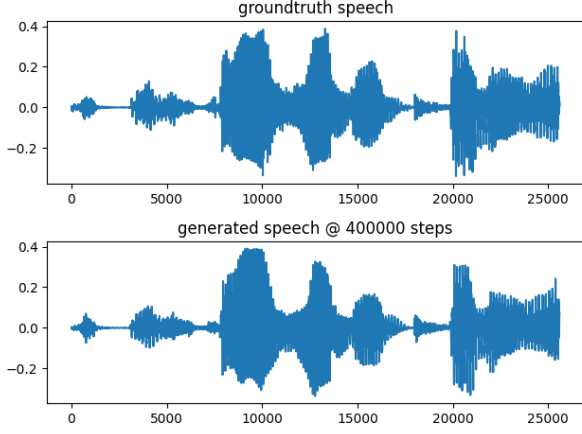


Figure 3: The plot for the audio generated by the ParallelWaveGAN vocoder after training at 400k steps versus the ground truth.

Table 3: The comparison of training epochs, training time and inference latency synthesis in different models. RTF (the real-time factor) denotes the time (in seconds) required for the system to synthesize one second waveform. The inference is done on NVIDIA GeForce GTX 1080 with 8GB and training on NVIDIA TITAN Xp GPU with 12GB

	Num epochs	Training time (h)	Inference speed (RTF)
Tacotron	200	49.58	0.12331
FastSpeech	1000	26.19	0.04349
FastSpeech2	1000	26.20	0.04355
Conformer	1000	66.71	0.04516

which sandwiches not only a multi-head self-attention module but also a convolution module. The convolution module itself has a pointwise convolution projecting into a glue activation layer, followed by 1-D depthwise convolution. Finally, there is swish activation, another pointwise convolution, and the batch norm. We trained FastSpeech, FastSpeech2, and Conformer for 500k steps.

We used ESPNet-TTS [23], which is an open-source end-to-end TTS toolkit that has implemented various state-of-the-art models. It provides recipes that are inspired by the Kaldi ASR toolkit. The training for all models was done on one NVIDIA TITAN Xp GPU with a batch size of 20. Table 3 depicts the comparison between different models in terms of the number of epochs, training time, and inference speed. Note that training time is only for the acoustic model, not including vocoder training. Except for Tacotron, the other models have the same inference speed. Conformer takes a longer time to train, but the quality is superior to other models. All of our pretrained models are free to download along with the original dataset.¹² We refer the readers to [23] for details of hyper-parameters and training configurations for each model.

5.3. Results

We randomly selected 30 fixed text samples with various lengths from the test corpus for evaluation. Different models

¹code:<https://github.com/roholazandie/ryan-tts>

²corpus:<http://mohammadmahoor.com/ryanspeech/>

Table 4: Comparison of MOS (mean opinion score) among our trained models with 95% confidence intervals.

model	MOS
Tacotron	3.00 ± 0.18
FastSpeech	3.27 ± 0.18
FastSpeech2	3.27 ± 0.18
Conformer	3.36 ± 0.17

all used the same text so that testers only reviewed the audio quality without interference factors. We then used a total of 120 audio samples generated by all four models for human evaluation. For a fair comparison, all the systems used Parallel WaveGAN as a vocoder. Twelve human subjects (native American English speakers, 18 years of age or older) participated in this study. Each participant independently scored all the models' outputs that were randomly shuffled. The evaluation was based on Likert scale score (1: Very Poor, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

Quantitative subjective evaluation, known as mean opinion score (MOS), has been used for results analysis [24]. MOS is simply the mean of the scores from all evaluators. Table 4 demonstrates the difference between the evaluations on each model. The best subjective result on our models is the conformer model. FastSpeech and FastSpeech2 have the same MOS score in our evaluations. Due to the effectiveness of variance information such as frequency and energy, as well as increased accuracy for predictions of duration, FastSpeech models are superior to Tacotron. However, the conformer model achieved the best score (closer to good on Likert scale) because of the ability of the model to harness the power of CNNs in order to extract local features in the creation of the Mel-spectrogram. Our results are comparable to results from the same models trained on LJSpeech [20].

To analyze the variance information, we calculated the first three moments of pitch distribution for the ground truth and synthesized speech [25, 26]. Table 2 shows the results. Among different models, Conformer is closer to the ground truth in terms of standard deviation, though Tacotron is closest with respect to skewness and kurtosis. This demonstrates that Conformer can produce pitch distribution close to natural voices, which results in better prosody.

6. Conclusions and Outlook

In this paper, we introduced the *RyanSpeech* corpus, which we designed primarily for TTS systems. We collected the text dataset from a variety of sources with a particular focus on conversational settings, and designed the corpus to be high-quality in its studio settings. As we have demonstrated, this is the first large, high-quality male speech corpus that is open-source. *RyanSpeech* can open up a number of possibilities for future research in the areas of ASR and TTS. With the trend in speech synthesis to find models that replicate human speech with small datasets, *RyanSpeech* proves an excellent candidate, especially for TTS evaluation and development in research as well as applications that require natural sources of speech, such as movies and podcasts.

7. Acknowledgements

Research reported in this paper was supported by the National Institute on Aging of the National Institutes of Health under award number R44AG059483 to DreamFace Tech, LLC.

8. References

- [1] K. Ito and L. Johnson, "The lj speech dataset," <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [2] 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," *arXiv preprint arXiv:1904.02882*, 2019.
- [3] 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)*. IEEE, 2015, pp. 5206–5210.
- [4] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Springer Science & Business Media, 2012, vol. 201.
- [5] J. Kominek, A. W. Black, and V. Ver, "Cmu arctic databases for speech synthesis," 2003.
- [6] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.
- [7] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [8] "Voxforge dataset," voxforge.org, free Speech Recognition,.
- [9] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Reviews*, 2014.
- [10] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.
- [11] A. Canavan, D. Graff, and G. Zipperlen, "Callhome american english speech," *Linguistic Data Consortium*, 1997.
- [12] Y.-J. Hu, C. Ding, L.-J. Liu, Z.-H. Ling, and L.-R. Dai, "The ustc system for blizzard challenge 2017," in *Proc. Blizzard Challenge Workshop*, 2017.
- [13] "M-ailabs," <https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/>, the m-ailabs speech dataset.
- [14] H. Abdollahi, M. H. Mahoor, R. Zandie, J. Siewierski, and S. H. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Transactions on Affective Computing*, p. Under review, 2021.
- [15] F. Dino, R. Zandie, H. Abdollahi, S. Schoeder, and M. Mahoor, "Delivering cognitive behavioral therapy using a conversational social robot," in *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on*, 2019.
- [16] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, D. Duckworth, S. Yavuz, B. Goodrich, A. Dubey, A. Cedilnik, and K.-Y. Kim, "Taskmaster-1: Toward a realistic and diverse dialog dataset," *arXiv preprint arXiv:1909.05358*, 2019.
- [17] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.1212303>
- [18] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.
- [19] 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.
- [20] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.
- [21] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.
- [22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [23] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7654–7658.
- [24] M. Chu and H. Peng, "Objective measure for estimating mean opinion score of synthesized speech," Apr. 4 2006, uS Patent 7,024,362.
- [25] B. Andreeva, G. Demenko, B. Möbius, F. Zimmerer, J. Jügler, and M. Oleskiewicz-Popiel, "Differences of pitch profiles in germanic and slavic languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [26] O. Niebuhr and R. Skarnitzl, "Measuring a speaker's acoustic correlates of pitch—but which? a contrastive analysis based on perceived speaker charisma," in *Proceedings of 19th International Congress of Phonetic Sciences*, 2019.