Narrator or Character: Voice Modulation in an Expressive Multi-speaker TTS

T Pavan Kalyan, Preeti Rao, Preethi Jyothi, Pushpak Bhattacharyya

IIT Bombay, India

190020124@iitb.ac.in, prao.ee.iitb.ac.in, pjyothi@cse.iitb.ac.in, pb@cse.iitb.ac.in

Abstract

Current Text-to-Speech (TTS) systems are trained on audiobook data and perform well in synthesizing read-style speech. In this work, we are interested in synthesizing audio stories as narrated to children. The storytelling style is more expressive and requires perceptible changes of voice across the narrator and story characters. To address these challenges, we present a new TTS corpus of English audio stories for children with 32.7 hours of speech by a single female speaker with a UK accent. We provide evidence of the salient differences in the suprasegmentals of the narrator and character utterances in the dataset, motivating the use of a multi-speaker TTS for our application. We use a fine-tuned BERT model to label each sentence as being spoken by a narrator or character that is subsequently used to condition the TTS output. Experiments show our new TTS system is superior in expressiveness in both A-B preference and MOS testing compared to reading-style TTS and single-speaker TTS.

Index Terms: Expressive TTS, speech synthesis, new TTS corpus, prosody modelling

1. Introduction

Text-to-speech systems target the acoustic realization of given text from its linguistic specification. Apart from the phone sequence, information about intonation, stress, and rhythm also influence the generated speech, ideally in a manner that represents the desired speaking style. Over the decades, research objectives have extended from achieving acceptable intelligibility and naturalness to expressiveness and other salient characteristics of the chosen speaking style. This has been accompanied by the shift from an explicit specification of suprasegmental and other high-level aspects to the implicit learning of the same from unannotated data. In such a situation, the quality of the training dataset plays a major role in the eventual performance of the TTS system. Well-known read speech datasets for TTS are LJSpeech [1], M-AILABS [2], and Blizzard 2013 [3]. Blizzard 2013 has multiple genres and has been used for single-speaker expressive TTS systems while Blizzard 2016 had a task involving a relatively small dataset from commercial audiobooks [4]. Story-telling to children is more interactive than simple reading aloud and is closer to conversational speech. However conversational speech datasets are not expressive enough [5]. In this work, we focus on the design of a corpus for a children's storytelling task, where modeling expressiveness, or prosody, is of prime importance.

Since manual annotation of training speech data for prosodic parameters is tedious, not the least due to the complexity of fully specifying prosody, it is more common to use architectures that learn latent representations directly from the ground truth speech audio [6, 7, 8, 9]. Given that recording new data of high quality for TTS training can be expensive, we try to identify available resources that can be adapted to our use case with relatively low effort. StoryNory[10] is a UK English podcast service with a human narrator reading out grade-appropriate stories to children in a naturally expressive style suited to the context. Each audio recording is accompanied by the story in text form. Apart from the narrated story, the audio recording contains extra utterances at the beginning (such as introductory remarks) and at the end (e.g. bidding goodbye). Finally, there is background music and animal sounds that typically occur during silences but occasionally overlap with the narrator's speech. In this work, we apply a suitable processing pipeline to convert this found data to a corpus[1] fit for training an end-to-end TTS system in the matching style. Further, a salient feature of the children's storytelling style is the use of voice modulation as the narrator switches to direct speech across different story characters in dialogs. We present and evaluate a method exploiting multi-speaker TTS as an effective way to emulate this.

End-to-end TTS architectures such as VAE [7] and GAN-based [8] models have been shown to produce high-quality speech using phoneme sequence and audio as input. Though many TTS models give output comparable to human speech, TTS models using GAN and Normalizing Flows [9] have performed better in expressiveness [11]. VITS [12] is a non-auto regressive TTS model using the Variational Auto-encoder architecture [13]. It employs normalizing flows [14] for modeling the prior distribution and GAN [15] pipeline to improve voice quality. VITS can be used in single-speaker and multi-speaker settings, where global conditioning similar to WaveNet [16] is used for multi-speaker TTS. We present A-B preference and MOS testing to evaluate our system for expressiveness with reference to a baseline model trained on a standard read-speech corpus. For more insights, we compute objective measures for intelligibility (ASR WER), as well as voice quality and prosodic parameter correlations with the ground-truth.

2. Dataset

The StoryNory[10] website has stories for children between the age of 7-11 years, narrated expressively by professional actors in UK accent, recorded in the studio. These stories are broadly classified into fairytales, classics by authors like Rudyard Kipling and Charles Dickens, and stories written by owners of the website. In this section, we discuss the processing pipeline used to obtain our TTS corpus and also present its salient features.

[1] https://github.com/tpavankalyan/Storynory (Code, labelled TTS dataset and checkpoints are avialable here.)

2.1. Data Segmentation and Filtering

The audio recording and accompanying story text are scraped from the website for the female narrator Natasha who has the single largest contribution of 66 hours of speech out of a total of 146 hours across 6 identified speakers. Modern TTS models require training data in the form of audio chunks of duration 10-20 seconds along with the matched transcript [1, 2, 3]. The scraped data is converted into this format by solving two problems: (i) obtaining the needed audio segments with aligned text and (ii) discarding noisy audio data (i.e. speech with background sounds).

The story text is split into sentences based on end-of-the sentence punctuations and quotation marks, with standard text normalization carried out. Audio segments corresponding to these sentences are obtained via Connectionist Temporal Classification (CTC) segmentation [17] similar to [18]. We use Nvidia-Nemo[2] to perform the CTC-segmentation that outputs a score indicating the probability of finding the correct alignment for each utterance. A poor score indicates poorly matched acoustics to the given text, such as that which arises with the extraneous speech in the beginning and the end. We use a threshold of –2 on this score to discard all such instances of mis-alignment. The above process reduced the audio data from 66 hours to 34.6 hours of properly aligned audio-text segments. To account for instances where the speaker might actually alter the reference text while speaking, we run these audio segments through QuartzNet ASR [19]. Inspecting the distribution of WER, we retain segments that have a WER of less than 10 %. This step also helped us eliminate speech segments with audible non-speech backgrounds. Thus, the final filtered dataset consists of 32 hours of audio data in the form of short utterances with the corresponding aligned transcripts as shown in Table 1.

Table 1: *Statistics of the Storytelling TTS dataset.*

Total duration	32.79 hours
Number of utterances	18641
Mean (s.d.) utterance duration	6.33 (5.04) sec
Total unique words	26148
Total distinct IPA phonemes	53
Total unique stories	251
Mean (s.d.) duration per character	0.09 (0.02) sec

2.2. Data Characteristics

Almost all stories present in this new dataset have an omniscient narrator. An omniscient narrator knows everything about the story, story event, character's thoughts, emotions and character events. Thus, in the audio, the narrator in the story switches between speaking the narrator sentences and character dialogues using voice modulation, and this was evident during informal listening. To understand the differences better in speech for narrator and character spoken sentences, we extracted multiple meaningful acoustic features for each type of speech utterance.

The first task is to identify the narrator and character spoken sentences from the story's text. The scraped dataset contained stories with punctuation and also stories with noisy punctuation. We use quotation marks to identify the character or narrator sentence. All the sentences of a story within the quotation marks are labelled as character sentences and the rest as narrator sentences. One example of noisy punctuation is an extra quotation mark, which can lead to incorrect labelling of a sentence as narrator or character. To avoid such instances, we collect sentences from stories having at least 10 character and narrator sentences, and have even number of quotation marks in the text. Since the TTS corpus was created after splitting the text based on quotation marks, we find these labelled sentences in the TTS corpus to identify the corresponding segmented audio files. Finally, 2000 audio files each for narrator and character voices were used to perform the acoustic analysis.

We consider two categories of acoustic features, viz. prosodic and voice-quality related. We hypothesize that the prosodic features (pitch, loudness, duration) are largely influenced by the syntax, semantics and emotion, while the voice quality changes most with assumed speaker identity (across the narrator and story characters). We used the OpenSmile [20] implementation of GeMAPS [21], a compact set of suprasegmental features found useful in speech emotion classification tasks. We find two kinds of aggregates for each utterance and acoustic parameter, namely the mean and variation. Both computed across the utterance, the mean is the average value and variation is the standard deviation representing contour dynamics, essential to expressivity. Table 2 show the average along with standard deviation of the utterance mean across the set of narrator and character utterances for those parameters that displayed statistically significant differences in both this mean value and its dynamics[3] (not shown in the table) across narrator and character classes (Mann-Whitney test, $p < 0.001$). We observe that the number of voice quality parameters discriminating character from narrator is relatively high, consistent with our hypothesis. The distribution differences motivate our design choice of separately modeling narrator and character-based utterances, further detailed in Section 3.

3. Methodology

Our storytelling TTS system is based on the VITS architecture. VITS [12] is a variational autoencoder-based TTS system that models the prior using normalizing flows and uses adversarial training for synthesising high-quality speech. The original formulation of VITS is shown in equation 1, where z is the latent space vector, c is an input phoneme sequence and $\log p_\theta(x|c)$ is the marginal likelihood of the data.

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z) - \frac{q_\phi(z|x)}{p_\theta(z|c)} \right] \quad (1)$$

The VITS architecture consists of a posterior encoder, a prior encoder, a decoder, a discriminator, and a stochastic duration predictor. The prior encoder models $p_\theta(z|c)$ and comprises a text encoder that converts input phonemes into hidden representations that are passed as input to a linear layer to predict the mean and variance of the prior distribution. Normalizing flows are used above this layer for improved modelling of the prior distribution. The hidden vectors are also passed to the stochastic duration predictor to estimate the distribution of phoneme durations. The text-encoder and stochastic duration predictor modules jointly predict the latent-space vectors that are extracted from the linear spectrogram using the posterior encoder ($q_\phi(z|x)$). The decoder ($p_\theta(x|z)$) then takes latent vectors as input and generates the raw speech waveform which is further input to the discriminator during training.

[2]https://github.com/NVIDIA/NeMo

[3]Refer to the supplementary material here: https://tinyurl.com/3mn56f85

Table 2: *Comparison of utt. mean (s.d.) of different acoustic features for narrator and character speech. The parameter names are from OpenSmile implementation of GeMAPS [21] with definitions given in the supplementary material.*

	Narrator	Character
Pitch		
F0 semitone RisingSlope	222.4 (201.1)	172.17 (208.91)
F0 semitone FallingSlope	67.99 (69.52)	66.56 (115.74)
Loudness		
loudness (in Sones)	0.50 (0.22)	0.70 (0.32)
loudness RisingSlope	9.58 (3.99)	11.22 (5.11)
loudness FallingSlope	7.45 (3.02)	9.15 (4.34)
Temporal		
Voiced duration (sec)	0.22 (0.11)	0.30 (0.20)
Unvoiced duration (sec)	0.19 (0.07)	0.20 (0.09)
Number of Syllables	15.07 (11.27)	10.47 (8.47)
Number of Pauses	1.24 (1.61)	0.75 (1.26)
Rate of speech (per sec)	2.67 (0.78)	2.53 (0.87)
Articulation rate (per sec)	4.06 (0.86)	3.82 (1.03)
Voice-quality		
alphaRatioV (in dB)	-14.38 (4.14)	-12.93 (4.68)
slopeV0-500	0.05 (0.04)	0.06 (0.04)
slopeV500-1500	-0.02 (0.01)	-0.01 (0.01)
logRelF0-H1-H2	7.15 (3.9)	7.73 (4.28)
logRelF0-H1-A3	21.00 (4.91)	18.69 (5.44)
HNRdBACF (in dB)	6.99 (1.95)	7.96 (2.08)
shimmerLocaldB (in dB)	1.15 (0.25)	1.07 (0.28)

The difference between the acoustic characteristics for the narrator and character roles, as discussed in Section 2.2, serves as our main motivation to provide these labels as additional conditional inputs to VITS. This new formulation can be realized by using VITS in multi-speaker setting. Multi-speaker VITS uses speaker embedding as additional conditional input to posterior encoder, decoder and prior encoder. The posterior encoder and prior encoder use WaveNet residual blocks and hence global conditioning is used to add the speaker embedding. The decoder uses an extra linear layer to transform the speaker embeddings that are added to the latent space vector predicted from the prior encoder. Similarly, the speaker embeddings added to the hidden vectors predicted from the text encoder are given as input to the stochastic duration predictor.

To accurately identify whether a given sentence is spoken by the narrator or a character, we introduced the narrator character (N-C) module, which is a pre-trained BERT model [22] with a 2-class classification layer. This model is fine-tuned using a dataset constructed from the new TTS corpus as mentioned in section 4.1. This module is useful when punctuations like quotation marks are missing or noisy in the input text. The output of this module is a label for narrator or character that can be input to Multi-speaker VITS as speaker label. Figure 1 shows the architecture of the multi-speaker VITS narrator character system.

4. Experiments and Results

We present analysis of the following systems:

1. VITS_SS: VITS trained on 32 hours of the StoryNory dataset with no modification to the VITS architecture.
2. VITS_LJS: VITS trained on the LJSpeech dataset.
3. VITS_NC: Our narrator character based VITS system. described in Section 3. As described in section 3, we modified VITS by adding the narrator character module to predict the speaker label, resulting in a system called VITS_NC.

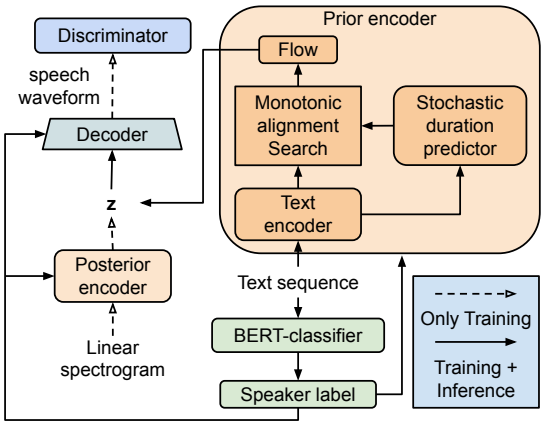

Figure 1: *Architecture of multi-speaker VITS with N-C module.*

4.1. Experimental Setup

The LJ Speech dataset is made up of 13,100 short audio clips from a single speaker. The total length of the clips is about 24 hours, and the audio format is 16-bit PCM with a 22.05 kHz sample rate. We used the dataset without making any modifications. Like VITS, we split the dataset randomly into three parts: a training set (12,500 samples), a validation set (100 samples), and a test set (500 samples). The StoryNory dataset is also sampled at 22.05 kHz and contains 18,640 audio clips from a single speaker. The total length of the clips is approximately 32 hours. The training, validation and test set contains 18,000, 100 and 540 samples respectively. Training proceeded similar to VITS, details of which are mentioned in the supplementary[3].

To train the N-C module, stories with at least 10 sentences within quotation marks were used as training data. The resulting dataset contains 14,173 sentences, with 8,905 labeled as narrator and 5,268 labeled as character dialogue based on the quotation marks. This data was randomly divided into a training set of 11,338 sentences and a validation set of 2,835 sentences. Punctuation was removed, and the Bert-base-cased model was utilized, with all layers frozen except for the classification layer. The N-C module achieved a validation accuracy of 93%. The fine-tuned model is used to label all sentences in the entire StoryNory dataset as either 0 (narrator) or 1 (character).

4.2. Subjective Evaluation

We conducted an A-B preference test with 25 L2 English-speaking participants to evaluate a single story selected from the test split of the dataset. Three types of sentences were chosen: sentences spoken only by the narrator, sentences spoken only by a character, and sentences that combine to contain a transition from narrator to character or vice versa. Each sentence conveyed different emotions such as happiness, suspense, anger, hurriedness, and fear. Participants were presented with two audio options for each sentence and asked to select the one that best matched the meaning and context of the sentence. The entire section of the story containing a particular sentence was shown as a reference to the participant. On average, each sentence consisted of 15 words. Two out of five systems were randomly selected for each participant,

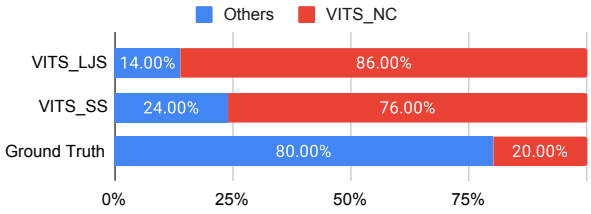

Figure 2: *Results of A-B preference test on expressiveness.*

and each system was played at least 5 times per sentence, with an average of 10 times. Figure 2 shows the results of A-B preference test where one of the systems is always fixed as VITS_NC. Listeners have preferred VITS_NC over VITS_LJS and VITS_SS more than 75% of the time.

For the Mean opinion score (MOS) test, we presented 20 sentences from the same story to a separate group of 30 listeners. Each system was judged at least 5 times by each listener, resulting in a total of 150 judgments per system [23]. The listeners were asked to rate the expressiveness of the generated audio on a scale of 1 (very poor) to 5 (excellent) as appropriate to the local story context. Table 3 shows the results of the MOS test for all the systems. VITS_NC outperforms VITS_LJS and VITS_SS and performs closer to the ground truth according to the MOS. These results imply that the multi-speaker VITS trained on Story Nory is better at synthesizing the storytelling speech as compared to single-speaker VITS trained on Story Nory or LJSpeech datasets. Low performing examples were found to arise mostly from narrator/character mislabeling.

Table 3: *Comparison of MOS (95% confidence intervals)*

Systems	MOS (CI)
Ground Truth (StoryNory)	3.96(±0.19)
VITS_LJS	3.03(±0.22)
VITS_SS	3.27(±0.18)
VITS_NC	3.62(±0.18)

4.3. Objective Evaluation

We used Whisper ASR [24] to evaluate the intelligibility of speech generated by each system. Table 4 contains the Word Error Rate (WER) and Character Error Rate (CER) of the speech samples generated for the test set. As the StoryNory dataset is more expressive, the output has more aspiration, larger variations in the speaker's intonation patterns and exaggerated stress compared to the LJSpeech dataset leading to higher WERs with the former. VITS_NC yields significantly lower WERs compared to VITS_SS, and is closer in WER to the ground-truth samples. This indicates that our multi-speaker VITS_NC model, that learns the acoustics of the narrator and character separately, results in more intelligible samples overall.

We also apply t-SNE analysis to the latent space of VITS_SS and VITS_NC to show the effect of multi-speaker training for the StoryNory dataset. We randomly sample 100 sentences each from the narrator and character sets and run the inference using VITS_NC and VITS_SS to obtain the latent vectors. Figure 3 shows the t-SNE plot for VITS_SS (left) and VITS_NC (right). The plots indicate the far superior clustering of the latent space into narrator and character for VITS_NC, validating the effectiveness of multi-speaker training.

Table 4: *Comparison of WER and CER of the test samples.*

Systems	WER %	CER %
Ground Truth (LJSpeech)	3.05	1.01
Ground Truth (StoryNory)	10.31	4.30
VITS_LJS	7.58	2.30
VITS_SS	16.55	6.86
VITS_NC	12.62	5.35

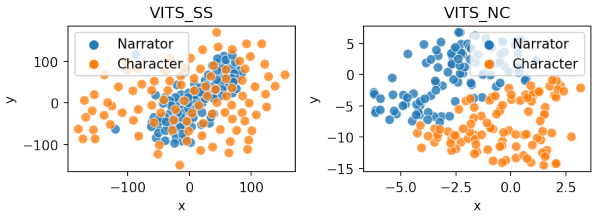

Figure 3: *t-SNE plot for latent space vectors obtained during inference from VITS_SS (left) and VITS_NC (right).*

Finally, we also estimate the Pearson correlation between the utterance-level acoustic features extracted from first 500 synthesized utterances of the test set and the corresponding Ground Truth (GT). Figure 4 shows the average over parameters within the same acoustic category as listed in table 2 that have statistically significant ($p < 0.01$) coefficient of correlation (r) for both GT-VITS_SS and GT-VITS_NC pairs. The r-values corresponding to the individual parameters are provided in the supplementary material[3]. VITS_LJS showed very low correlation values and is not included. We note that features from VITS_NC are indeed better matched to the ground-truth with especially prominent improvements over VITS_SS for temporal and certain voice-quality features.

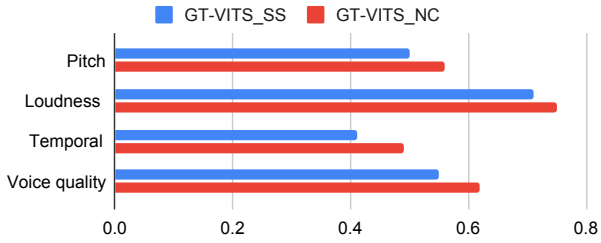

Figure 4: *Pearson correlation coeffecient for GT-VITS_SS and GT-VITS_NC pairs averaged over aggregates of parameters from the same acoustic group mentioned in table 2.*

5. Conclusions

Using a suitable data preparation pipeline, we present the most extensive single-speaker storytelling TTS corpus out of available audio podcasts. Our prosodic analysis indicates that storytelling to children requires voice modulation on top of single-speaker expressiveness. We investigated the conditioning of the VITS model with automatically detected speaker identity (narrator or character) to obtain significant improvements in expressiveness and voice quality. Future work could involve generating more fine-grained distinctions between different character voices based on textual information drawn from the story text.

6. References

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

[2] M. A. I. L. GmbH, "The M-AILABS speech dataset," https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/, 2019, accessed: 2023-01-12.

[3] S. King and V. Karaiskos, "The Blizzard Challenge 2013," 2014.

[4] S. Raptis, P. Tsiakoulis, A. Chalamandaris, and S. Karabetsos, "Expressive speech synthesis for storytelling: the innoetics' entry to the blizzard challenge 2016," in *Proc. Blizzard Challenge*, 2016.

[5] R. Milner, M. A. Jalal, R. W. Ng, and T. Hain, "A cross-corpus study on speech emotion recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 304–311.

[6] Y. Wang, R. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[7] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6945–6949.

[8] ShuangMa, D. McDuff, and Y. Song, "Neural TTS stylization with adversarial and collaborative games," in *International Conference on Learning Representations (ICLR)*, April 2019. [Online]. Available: https://www.microsoft.com/en-us/research/publication/neural-tts-stylization-with-adversarial-and-collaborative-games/

[9] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, "Using Vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6179–6183.

[10] H. Fraser. (2023) Storynory. [Online]. Available: https://www.storynory.com/

[11] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting over-smoothness in text to speech," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8197–8213. [Online]. Available: https://aclanthology.org/2022.acl-long.564

[12] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[14] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International Conference on Machine Learning*, 2015.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[16] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*.

[17] L. Kurzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-Segmentation of large corpora for german end-to-end speech recognition," in *International Conference on Speech and Computer*, 2020.

[18] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi multi-speaker english TTS dataset," in *Interspeech*, 2021.

[19] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6124–6128, 2019.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[23] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? no! - an empirically-supported critique of interspeech 2014 TTS evaluations," in *Interspeech*, 2015.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, vol. abs/2212.04356, 2022.