Towards Natural-Sounding Text to Speech in English

Kriss Saulitis[0009−0008−2730−0544], Evalds Urtans[0000−0001−9813−0548], and Vairis Caune[0000−0002−0893−8998]

Faculty of Information Technologies, Ventspils University of Applied Sciences, Inzenieru iela 101, Ventspils, LV-3601, Latvia
kriss@saulitis.dev,
evalds.urtans@venta.lv,
vairis.caune@venta.lv

Abstract. This study focuses on a systematic review of the literature and an experimental comparison of 20 English speech synthesis methods. Nine of the models were subjected to a quantitative analysis, using selected samples from the Common Voice data set and using criteria to assess both quality and precision. The research methodology includes the configuration of speech synthesis models to generate audio samples, which are then used to compare models based on established criteria. The NISQA model is used to evaluate speech quality through machine learning, mimicking the subjective MOS metric. Character and word error rate metrics are used to evaluate the precision of the synthesized samples. The CoMoSpeech model showed the best quality indicators (MOS - 3.85), while the VITS model demonstrated the highest precision (CER - 1.48%) and the total average of the metric. [1]

Keywords: TTS · survey · speech synthesis · deep-learning · diphone concatenation · literature analysis · NISQA · objective comparison

1 Introduction

The field of speech synthesis is rapidly expanding. It has seen significant growth in research and technological advancements. This assertion is corroborated by a systematic query conducted on the Semantic Scholar database, revealing that, over the course of a decade, the field has witnessed the publication of 3,060 scholarly articles. In addition, 475 of these publications have emerged in the last year. Speech synthesis technology, which aims to replicate human voice through computerized methods, finds extensive application in various sectors, underscoring its importance alongside advances in artificial intelligence and machine learning, particularly deep learning and neural networks Mehrish et al. (2023). This study focuses on a systematic review and comparison of 20 notable English language

[1] Synthesized samples and code repository are available at https://research.saulitis.dev/english-speech-synthesis-comparison-2024

speech synthesis models, along with an objective analysis using selected samples from the Common Voice data set and quality criteria to further analyze 9 models. Performing this requires the use of the NISQA machine learning model [2] and the Asya STT model [3]. By providing an objective comparison of models that have openly accessible weights and code, this research helps identify the best models for future improvements or training in additional languages.

2 Related work

The advancement of speech synthesis has been significantly influenced by two primary methodologies: diphone concatenation and deep machine learning techniques. The diphone concatenation method uses speech segments called diphones to synthesize speech. These are linked at their stable midpoints using the Multi-Band Resynthesis OverLap Add (MBROLA) system to minimize discontinuities and enhance intelligibility. However, this method can produce speech that sounds unnatural or is overly synthesized. D'Alessandro et al. (2005)

In contrast, the deep machine learning method uses various types of neural-based TTS network. Acoustic models are used to generate acoustic features from linguistic features, phonemes, or graphemes. Tan (2023) To achieve this, different model structures have been adopted, such as Recurrent Neural Networks (RNNs) Wang et al. (2017), Convolutional Neural Networks (CNNs) Arık et al. (2017) or Transformers Li et al. (2019). For more advanced spectrogram generation Generative Adversarial Networks (GANs), Flow, Variational Autoencoders VAE or Diffusion models can be used. Vocoder models, on the other hand, generate waveforms from given acoustic features and are often paired with an acoustic model that generates these acoustic features. For generation autoregressive architectures Oord et al. (2016), flow-based architectures Kim et al. (2020), GAN architectures Donahue et al. (2018) and diffusion architectures Chen et al. (2020) can be used. However, in recent years, fully end-to-end based TTS models have emerged that combine acoustic feature generation and vocoder parts in a single model in order to simplify the training process and speed up inference. One of the first models to do so was FastSpeech 2 Ren et al. (2020), which synthesized speech directly from a given text.

For evaluation and comparison of the methods, two primary metric groups are used: precision, which includes the Word Error Rate (WER) and the Character Error Rate (CER), and quality, which includes the Mean Opinion Score (MOS), which is prevalent in telecommunications and speech synthesis, evaluates audio quality based on subjective judgments of humans on a Likert scale from 1 (low quality) to 5 (high quality). This metric can be adapted to evaluate other features such as naturalness, noise, or volume. Shen et al. (2023)

[2] NISQA repository - `https://github.com/gabrielmittag/NISQA`
[3] `https://explorer.asya.ai`

3 Systematic literature review

The search for speech synthesis methods and their scientific literature used tools, techniques, and databases, including "Papers with Code", Semantic Scholar, IEEE, ArXiv, and Google Scholar. The key phrases used were: "TTS methods", "TTS model comparison", "Speech synthesis models", "SOTA TTS models", and "Diphone-based TTS systems". Furthermore, references within the original work led to the identification of other models related to speech synthesis. A subsequent reference search was conducted using Google Scholar. Models dated before 2021 required at least 100 citations. A total of 20 speech synthesis methods were selected for the detailed analysis and evaluation of established quality criteria. An overview of the methods can be seen in Figure 1.

Tacatron-based models include Tacotron 2 Shen et al. (2018), FastSpeech 2 Ren et al. (2020), and JETS Lim et al. (2022), where each improved on previous work. Flow-based models were popularized by Glow-TTS Kim et al. (2020), which built the foundation for other flow-based models, such as OverFlow Mehta et al. (2022), YourTTS Casanova et al. (2022), VITS Kim et al. (2021) and VITS 2 Kong et al. (2023). Diffusion-based models include Grad-TTS Popov et al. (2021), Guided-TTS 2 Kim et al. (2022) and CoMoSpeech Ye et al. (2023), where the last is distilled Grad-TTS. The models of the newer Vector Quantization-Based Method include DelightfulTTS 2 Liu et al. (2022), FoundationTTS Xue et al. (2023), MQTTS Chen et al. (2023), VQTTS Du et al. (2022). NaturalSpeech Tan et al. (2022) started with the flow-based method but was later adopted to include the vector quantization method in NaturalSpeech2 Shen et al. (2023). For older phoneme-based models, eSpeak Duddington (1995) and MaryTTS Schröder and Trouvain (2003) were selected.

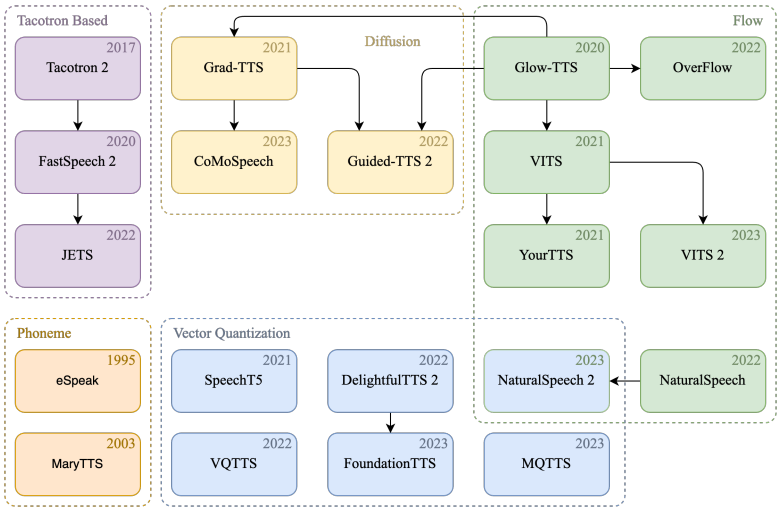

Fig. 1: Historical overview of Speech synthesizer methods

For each method, original quality and precision metrics were also compiled and processed. The results can be seen in Table 1, where only metrics and data sets that were comparable with at least one other method were added.

Table 1: Quality metrics for datasets provided in papers

Nr	Title	LJSpeech	VCTK	LibriTTS
1	**CoMoSpeech**	4.24		
2	**VITS 2**	4.47	3.99	
6	**OverFlow**	3.43		
8	**Guided-TTS 2**	4.23	4.23	
9	**NaturalSpeech**	4.56		
10	**VQTTS**	**4.71**		
11	**JETS**	4.02		
12	**YourTTS**		4.24	**4.25**
13	**SpeechT5**			3.65
14	**VITS**	4.43	**4.38**	
15	**Grad-TTS**	4.44		
16	**FastSpeech 2**	3.83		
17	**Glow-TTS**	4.01		3.45

The following quality criteria were applied to filter the selected works, aiding in narrowing down the scope for further method comparison:

1. The training of the model utilized a data set that is publicly accessible;
2. The method has available synthesized speech examples;
3. The model's source code is publicly available and completely reproduces the model of the published paper;
4. The model has freely available training result weights;
5. At least one of the methods should rely on the diphone concatenation method.
6. Existing published samples should be of comparable quality to those produced by the locally configured method.

For the diphone concatenation method, the MaryTTS system was used for further comparisons, as it is more recent and produces significantly better results than eSpeak. By including this type of method, we can further compare the impact and quality improvements provided by machine learning methods. Data for quality criteria is scored and compared in Table 2.

Table 2: Model and system quality metric results

Nr	Title	K1	K2	K3	K4	K5	K6	Score
1	**CoMoSpeech**	+	+	+	+		+	**5**
2	**VITS 2**	+	+	+				3
3	**NaturalSpeech 2**	+	+	+				3
4	**FoundationTTS**	+						1
5	**MQTTS**	+	+	+	+		+	**5**
6	**OverFlow**	+	+	+	+		+	**5**
7	**DelightfulTTS 2**	+	+					2
8	**Guided-TTS 2**	+	+	+				3
9	**NaturalSpeech**	+	+					2
10	**VQTTS**	+	+					2
11	**JETS**	+	+	+	+			4
12	**YourTTS**	+	+	+	+		+	**5**
13	**SpeechT5**	+		+	+		+	4
14	**VITS**	+	+	+	+		+	**5**
15	**Grad-TTS**	+	+	+	+		+	**5**
16	**FastSpeech 2**	+	+	+	+		+	**5**
17	**Glow-TTS**	+	+	+	+		+	**5**
18	**Tacotron 2**		+	+	+		+	4
19	***MaryTTS***					+	+	**2**
20	**eSpeak**					+	+	2

4 Methodology

For conducting an objective comparison of speech synthesis methods, the following steps were undertaken:

1. Selection and study of method publications, using selection and quality criteria.
2. Model configuration for speech sample synthesis, ensuring that they are of similar quality to the samples provided by the authors.
3. Synthesis of speech samples and their processing using APIs of the ASR (Automatic Speech Recognition) system APIs.
4. Calculation of quality indicators using the NISQA model; and accuracy calculation using CER (Character Error Rate) and WER (Word Error Rate) metrics, based on data obtained from the ASR system.

4.1 Data Sets

To obtain validation data sets for each of the speech synthesis methods and models, it is necessary to acquire a uniform textual data set to generate audio files. Several criteria were established for the selection of such data sets to ensure sufficient text diversity and sample count to obtain objective results.

The main criteria established for such a data set are:

1. The data set has not been used in the training of any models;
2. Synthesis of the data set examples into audio files results in speech not longer than 10 seconds;
3. The data set contains at least 4,000 examples with high diversity to conduct a comprehensive analysis of method quality;
4. The data set's license permits its use in scientific research.

Initially, data sets used by the selected speech synthesis methods were compiled. This compilation is shown in Table 3. Parameters of the most popular data sets were also compiled from model descriptions:

- **LJSpeech** - a single-speaker English language data set containing approximately 13,100 audio samples.
- **VCTK** (*Voice Cloning Toolkit*) - an English language data set with various accents, containing 109 different speakers and approximately 44,000 audio samples.
- **LibriSpeech** - an English language data set containing 2,484 different speakers and approximately 1,000 hours of recorded text.
- **LibriTTS** - an English language data set containing 2,456 different speakers and approximately 1,000 hours of recorded text.

Subsequently, a search for other data sets was performed, resulting in the selection of the "*Mozilla Common Voice*" data set. This data set covers text-audio pairs in various languages. For this study, the English data set, consisting of 1,752,390 recordings, was chosen. To obtain higher-quality data and narrow down the volume of synthesized and tested files, they were filtered using the following criteria:

1. At least 7 reviewers have positively evaluated the audio recording;
2. No more than 10% of reviewers have negatively evaluated the audio recording;
3. Recordings labeled as "*Benchmark*" were excluded;
4. In case of duplicates, the recording with the highest number of positive evaluations was selected.

As a result, 4,677 recordings were obtained, which were then used to synthesize the model test sets.

Table 3: Model datasets

Nr	Title	Datasets
1	**CoMoSpeech**	LJSpeech
2	**VITS 2**	LJSpeech; VCTK
3	**EfficientSpeech**	LJSpeech
4	**NaturalSpeech 2**	VCTK; LibriSpeech
5	**FoundationTTS**	VCTK; LibriTTS; Proprietary
6	**MQTTS**	VoxCeleb; GigaSpeech
7	**OverFlow**	LJSpeech
8	**Estonian TTS**	Proprietary
9	**DelightfulTTS 2**	Proprietary
10	**Guided-TTS 2**	LJSpeech; VCTK; LibriSpeech; LibriTTS; VoxCeleb; LibriLight
11	**NaturalSpeech**	LJSpeech; Proprietary
12	**VQTTS**	LJSpeech
13	**JETS**	LJSpeech
14	**YourTTS**	VCTK; LibriTTS; TTS-Portuguese; M-AILABS; MLS
15	**EdiTTS**	LJSpeech
16	**SpeechT5**	LibriSpeech; LibriTTS
17	**VITS**	LJSpeech; VCTK
18	**Grad-TTS**	LJSpeech
19	**Parallel Tacotron 2**	Proprietary
20	**Apple-TTS**	Proprietary
21	**FastSpeech 2**	LJSpeech
22	**Glow-TTS**	LJSpeech; LibriTTS
23	**Tacotron 2**	Proprietary
24	**MaryTTS**	Proprietary

4.2 Synthesis of Test Sets

Based on the quality criteria established for the methods, 9 selected methods were configured locally. The configuration was based on publicly available source code and publicly available weights. If a model supported the synthesis of different voices, the first voice synthesis embedding vector recommended by the authors was chosen. Each configured method was run in inference mode using a test set and resulted in a synthesized audio data set that was used to calculate the metrics of each model.

4.3 Metrics

To objectively assess the accuracy of speech synthesis, WER (Word Error Rate) and CER (Character Error Rate) metrics were used, which required conversion of test audio data into text using the Asya.ai ASR system, followed by metric calculation for each recording and an average for the speech synthesis method, with results consolidated in Table 5. Furthermore, the NISQA model was used to evaluate speech quality through machine learning, evaluating overall quality

(Mimicking MOS), naturalness, coloration, noise, discontinuity, and loudness, with higher scores indicating superior speech quality. These evaluations were carried out on the synthesized speech test dataset, with average results for each metric per model shown in Table 4, providing comprehensive information on both the precision and quality of synthesized speech.

5 Results

In total, 20 methods were reviewed, of which 9 met established quality criteria and were selected. Each method was installed in a local environment, and audio test sets were synthesized.

Table 4: System and data set NISQA metrics

Nr	Title	Quality	Naturalness	Coloration	Noisiness	Discontinuity	Loudness
1	**CoMoSpeech**	**3.85**	4.41	**4.41**	**4.67**	**4.58**	3.97
5	**MQTTS**	3.54	**4.53**	4.24	4.61	4.33	**4.12**
6	**OverFlow**	3.08	3.93	4.17	4.50	4.02	3.55
12	**YourTTS**	3.24	4.02	3.73	4.30	4.09	4.04
14	**VITS**	3.07	4.29	4.10	4.43	4.27	3.66
15	**Grad-TTS**	3.64	4.36	4.38	4.65	4.57	3.89
16	**FastSpeech 2**	2.87	3.44	3.72	4.04	3.62	3.40
17	**Glow-TTS**	3.04	3.85	4.16	4.52	4.03	3.78
19	***MaryTTS***	*2.38*	*3.35*	*3.43*	*3.17*	*3.73*	*3.72*
	Common Voice	*3.25*	*3.38*	*3.36*	*3.87*	*3.76*	*3.52*

The NISQA metrics can be seen in Table 4. CoMoSpeech model achieved the highest scores, on average 22% better (see Figure 2) than the chosen data set. High ratings were also obtained by the MQTTS and Grad-TTS models. Interestingly, the chosen data set (*Common Voice*) speech samples set one of the lowest results, despite being selected samples with the highest ratings. It is very likely that the results are as such precisely because this data set was recorded by people in home conditions, with various quality microphones and background noise. The worst rating, averaging 7%, was obtained using the MaryTTS diphone concatenation-based method, which sounded overly robotic. The most iconic examples based on this indicator are compiled on a publicly available website. [4]

CER and WER metrics are summarized in Table 5. VITS model synthesized text more precisely, on average 30% better (see Figure 2) than the chosen data set (*Common Voice*) speech samples for the same metric. The other methods, except MaryTTS, which was similar to the baseline, achieved worse results in this metric. Notable examples of these indicators are also available on a publicly available website. [4]

[4] `https://research.saulitis.dev/english-speech-synthesis-comparison-2024`

Table 5: System and data set CER, WER metrics

Nr	Title	WER	CER
1	**CoMoSpeech**	0.07	0.03
5	**MQTTS**	0.21	0.18
6	**OverFlow**	0.05	0.02
12	**YourTTS**	0.11	0.05
14	**VITS**	**0.04**	**0.01**
15	**Grad-TTS**	0.06	0.03
16	**FastSpeech 2**	0.06	0.02
17	**Glow-TTS**	0.08	0.04
19	***MaryTTS***	*0.05*	*0.02*
	Common Voice	*0.05*	*0.02*

Overall, by calculating the average results for both groups of metrics, the VITS speech synthesis model was in the lead, being very precise and also obtaining a high rating in terms of speech quality. The next best model could be identified as *Grad-TTS*, which reached a slightly lower average value than the *Common Voice* data set.

All source codes for the speech processing and synthesis algorithm are available in the project repository [5]. There, one can also find the unprocessed versions of all the model data and indicators.

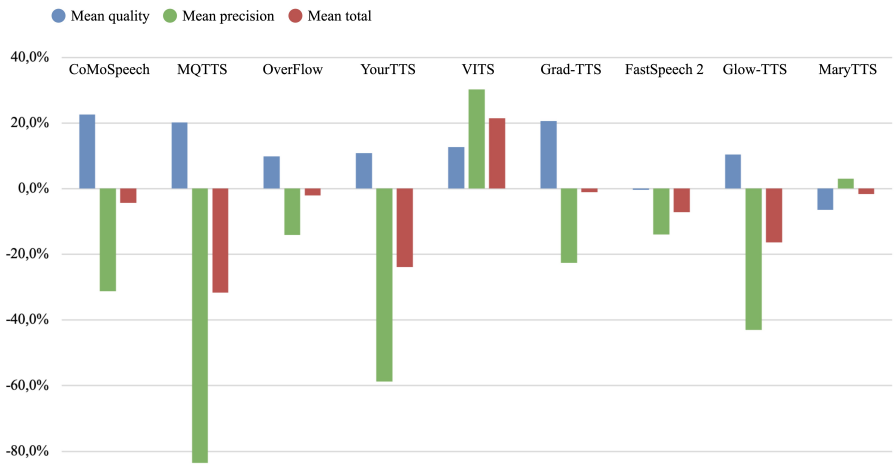

Fig. 2: Mean quality, accuracy, and total scores against Common Voice results as a baseline

[5] https://github.com/krsaulitis/tts-survey-2023

6 Conclusions

This study aimed to review and analyze speech synthesis methods, particularly those in the English language, to offer a clear and objective comparison between them. We concluded that comparing models is inherently challenging due to the variance in dataset usage and the reliance on non-objective metrics. However, this challenge can be mitigated by adopting objective precision and quality metrics such as CER, WER, and NISQA, which have been shown to be reliable indicators. Among the models evaluated, the CoMoSpeech model emerged as the superior in terms of quality, achieving a Mean Opinion Score (MOS) of 3.85. In contrast, the VITS model was identified as the most accurate, with a Character Error Rate (CER) of 1.48% and having the highest average overall score. This underlines the importance of evaluating both the quality and accuracy of speech synthesis methods, since these metrics do not always align, highlighting the intricate balance required to optimize both aspects in the development of speech synthesis technologies.

Bibliography

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., et al. (2017). Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*, pages 195–204. PMLR.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. (2022). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Chen, L.-W., Watanabe, S., and Rudnicky, A. I. (2023). A vector quantized approach for text to speech synthesis on real-world spontaneous speech. *ArXiv*, abs/2302.04215.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. (2020). Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

D'Alessandro, N., Sebbe, R., Bozkurt, B., and Dutoit, T. (2005). Maxmbrola: A max/msp mbrola-based tool for real-time voice synthesis. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE.

Donahue, C., McAuley, J., and Puckette, M. (2018). Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*.

Du, C., Guo, Y., Chen, X., and Yu, K. (2022). Vqtts: High-fidelity text-to-speech synthesis with self-supervised vq acoustic feature. *ArXiv*, abs/2204.00768.

Duddington, J. (1995). espeak text to speech.

Kim, J., Kim, S., Kong, J., and Yoon, S. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Kim, S., Kim, H., and Yoon, S.-H. (2022). Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data. *ArXiv*, abs/2205.15370.

Kong, J., Park, J., Kim, B., Kim, J., Kong, D., and Kim, S. (2023). Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. *arXiv preprint arXiv:2307.16430*.

Li, N., Liu, S., Liu, Y., Zhao, S., and Liu, M. (2019). Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.

Lim, D., Jung, S., and Kim, E. (2022). Jets: Jointly training fastspeech2 and hifi-gan for end to end text to speech. pages 21–25.

Liu, Y., Xue, R., He, L., Tan, X., and Zhao, S. (2022). Delightfultts 2: End-to-end speech synthesis with adversarial vector-quantized auto-encoders. In *Interspeech*.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., and Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, page 101869.

Mehta, S., Kirkland, A., Lameris, H., Beskow, J., Székely, É., and Henter, G. E. (2022). Overflow: Putting flows on top of neural transducers for better tts. *arXiv preprint arXiv:2211.06892*.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. (2021). Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *ArXiv*, abs/2006.04558.

Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., and Bian, J. (2023). Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *ArXiv*, abs/2304.09116.

Tan, X. (2023). *Neural text-to-speech synthesis*. Springer Nature.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F. K., Qin, T., Zhao, S., and Liu, T.-Y. (2022). Naturalspeech: End-to-end text to speech synthesis with human-level quality. *ArXiv*, abs/2205.04421.

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

Xue, R., Liu, Y., He, L., Tan, X., Liu, L., Lin, E., and Zhao, S. (2023). Foundationtts: Text-to-speech for asr custmization with generative language model. *arXiv preprint arXiv:2303.02939*.

Ye, Z., Xue, W., Tan, X., Chen, J., fei Liu, Q., and Guo, Y.-T. (2023). Comospeech: One-step speech and singing voice synthesis via consistency model. *Proceedings of the 31st ACM International Conference on Multimedia*.