Interspeech 2024
1-5 September 2024, Kos, Greece

Exploring the Robustness of Text-to-Speech Synthesis Based on Diffusion Probabilistic Models to Heavily Noisy Transcriptions

Jingyi Feng, Yusuke Yasuda, Tomoki Toda

Nagoya University, Japan

feng.jingyi@g.sp.m.is.nagoya-u.ac.jp, yasuda.yusuke@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

Large data volumes can benefit text-to-speech (TTS), but speech data with high-quality annotation is limited. Automatic transcription enables the transcription of found speech data to enhance the data volume for TTS, but TTS training suffers from transcription errors. In this paper, we investigate the robustness of typical TTS models against heavily noisy transcripts, including diffusion, flow, and autoregressive-based TTS models, in terms of objective intelligibility and subjective naturalness. Our experimental results show that diffusion-based TTS is extremely robust to heavily noisy transcriptions, mitigating about 30% of the word error rate compared to autoregressive and flow-based models. We also show that iterative inference with a long diffusion time is key to the robustness of diffusion-based TTS based on likelihood analysis.

Index Terms: speech synthesis, noisy transcription, diffusion-based model

1. Introduction

Utilizing the powerful learning capabilities of deep neural networks, current text-to-speech (TTS) methods are capable of generating high-quality speech that borders on human naturalness [1, 2, 3, 4, 5]. Most of these models are trained on high-quality annotated data, but collecting such data is difficult. To achieve TTS robust to pronunciation errors by learning abundant texts, the utilization of unannotated data resources is promising. However, unannotated speech data often lack accurate transcriptions, which makes it critical to train TTS models stably.

Automatic speech recognition (ASR) is a straightforward method to transcribe unannotated speech data automatically. The automatic transcriptions can be used as training data for TTS. This approach has to face the problem of transcription errors of ASR leading to a mismatch between transcription and speech, and they are usually considered to affect the quality of synthesized speech seriously [6]. In study [7], the robustness of an end-to-end TTS model based on the attention mechanism on different noisy transcriptions was investigated by training on specific error types and ASR transcription error data. The attention mechanism in TTS shows some robustness against ”insertion” errors, but the model is severely compromised by ”deletion” and ”substitution” errors.

Recently, diffusion-based TTS has been reported to have robustness to noise in linguistic inputs [8]. In the study, diffusion-based TTS shows robustness for linguistic inputs with ambiguous orthography. Since diffusion-based TTS can perform unconditional and classifier-guided generation without transcripts [9] and infilling of incomplete speech [10], it is promising to handle noisy transcriptions produced by ASR.

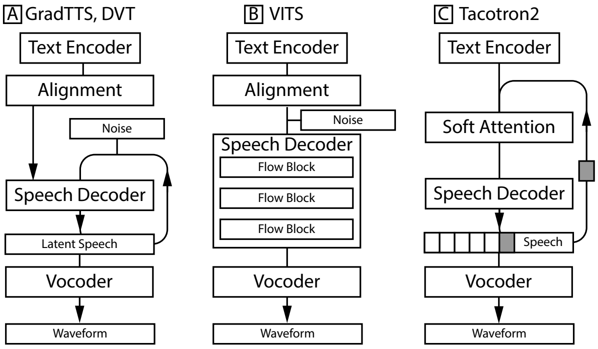

Figure 1: *TTS Architectures*

This paper aims to probe more deeply into the robustness of diffusion-based TTS methods, along with mainstream TTS methods, in the face of noisy transcriptions. We target Tacotron2[1], VITS[5], GradTTS[11], and DVT[8] as TTS methods. Among them, GradTTS and DVT are diffusion-based TTS methods. We perform ASR on different levels of noisy speech to obtain transcription data with different levels of impairment and apply these to the training of TTS models. In the study, we evaluate the robustness of the models in terms of the intelligibility and naturalness of the synthesized speech.

Our contributions are summarized as follows: (1) We train diffusion, flow, and autoregressive-based TTS models using transcripts with various noise levels and reveal that diffusion-based TTS methods show high robustness to transcription noise; (2) We analyze the inference process of diffusion with likelihood ratios and show that training and inference with extended diffusion time realizes the high robustness.

2. Transcription Noise Robustness of TTS

In TTS, three components can potentially handle noisy transcripts: text encoder, alignment, and decoder with a probabilistic model. We mainly focus on the decoder with a diffusion probabilistic model.

2.1. Diffusion probabilistic models

Figure 1A shows the architecture of diffusion-based TTS. The diffusion probabilistic model generates speech $\boldsymbol{x}_0$ from noise $\boldsymbol{x}_T$ via many latent speech $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$, where $t = 1, 2, \ldots, T$ is diffusion time. The diffusion-based TTS models the probability of latent speech $\boldsymbol{x}_{t-1}$ given the previously generated latent speech $\boldsymbol{x}_t$ and linguistic input $\boldsymbol{y}$ as $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})$. Therefore, the model can use noisy speech $\boldsymbol{x}_t$ and linguistic informa-

10.21437/Interspeech.2024-2337

tion $\boldsymbol{y}$ during inference. This mechanism lets the model gradually decode high-quality speech from both latent speech and transcripts via iterative denoising. During iterative inference, diffusion-based TTS transitions its behavior. The latent speech is not informative during the early inference because its noise level is too high. On the other hand, latent speech is informative during the late inference stage because it becomes close to target speech. The linguistic inputs $\boldsymbol{y}$ are reliable during whole inference in normal TTS. Therefore, the diffusion-based TTS model is expected to depend on linguistic inputs more than noisy latent speech during the early inference stage and on latent speech more than linguistic inputs during the late inference stage. This expectation of diffusion-based TTS is observed in normal TTS where transcripts are correct [8]. As a result, diffusion-based TTS shows robustness to ambiguous orthographies, such as characters in the English language, decoding correct pronunciation gradually from latent speech and transcripts via iterative denoising.

The behavior of diffusion-based TTS is not investigated when linguistic inputs $\boldsymbol{y}$ are partially unreliable due to annotation or automatic transcription errors. Intuitively, diffusion-based TTS may also be robust to transcription errors because it can decode correct pronunciation gradually from latent speech with less dependence on transcripts via iterative denoising. These characteristics of diffusion-based TTS suggest that it may generate natural speech from heavily corrupted transcripts.

2.2. Generative flow and autoregressive models

Figure 1B shows the architecture of flow-based TTS. Unlike diffusion and autoregressive models, flow-based TTS uses single-step inference to predict whole speech sequences. Internally, it consists of compositions of many flow-block functions with denoising features. The flow-based decoder decodes speech from noise via many latent variables by applying flow-block functions. Thanks to these mechanisms, flow-based models can overcome the drawbacks of autoregressive models. Based on the partial similarities between flow and diffusion models, flow-based TTS may have the robustness to noise in transcripts as diffusion-based TTS. The behavior of flow-based TTS under heavily corrupted transcripts has not been fully investigated.

Figure 1C shows the architecture of autoregressive TTS. The autoregressive model is one of the major probabilistic models used for TTS. It can generate highly natural speech as represented by Tacotron2 [1]. It has an iterative inference mechanism as diffusion models. However, it predicts a speech frame at each iteration instead of the whole latent speech sequence, and its inference does not involve denoising speech via many latent variables. Therefore, it has drawbacks of unstable and slow inference. Because its inference depends on the previous output of the speech frame, poor speech frames are generated once an unnatural speech frame is predicted at some point. These characteristics of the autoregressive model indicate it has only limited robustness to noisy transcriptions.

2.3. Alignments

An alignment module in TTS is a component that relates text to speech. There are two major alignment methods for TTS: soft attention [12, 13, 14, 15] and hard alignment [16, 17, 18]. Soft attention represents alignments as weights over text inputs by attending to relevant parts of inputs for corresponding speech output. The soft representation of alignments in the soft attention allows skipping irrelevant parts of the input sequence. This feature enables TTS models with soft attention to avoid errors by skipping over ”insertion” errors in noisy transcriptions. Thus, this type of model shows a certain degree of robustness in the face of noisy transcriptions [7]. However, this robustness is expected to be limited because soft attention can do nothing against ”deletion” and ”substitution” errors. A more severe drawback of soft attention is fatal alignment errors. Because TTS should read texts from left to right, alignments between text and speech must be monotonic. However, soft attention can not guarantee the monotonic structure of alignments, which may cause fatal alignment errors such as skip, repeat, and termination. To circumvent alignment errors, soft attention is enhanced with location features [15, 1].

On the other hand, hard alignment attends to only a single input token for an output of speech frame. Thanks to its discrete nature, hard alignment can guarantee monotonic alignments. In hard alignments, dynamic programming is used to derive monotonic alignments. The hard alignments can be represented in phoneme duration units for efficient and robust inference. The monotonicity and duration-level representation of hard alignments are expected to provide robustness to noisy transcripts.

2.4. Text encoder

The text encoder encodes linguistic inputs into acoustic representations. The network structure of the text encoder affects the performance of TTS. For example, Tacotron2 uses a simpler CNN-based encoder for phoneme inputs. In contrast, Tacotron [19] uses a complex CBHG encoder to handle character inputs inspired by character-aware word embedding [20]. It is shown that the CBHG encoder performs better than the CNN-based encoder for characters and small models[21]. Accordingly, the encoder structure is expected to affect the robustness of transcription noise.

2.5. TTS architectures

TTS architectures are a combination of decoder, alignment, and text encoder components. We investigate GradTTS [11] and DVT [8] as diffusion-based TTS, VITS as flow-based TTS, and Tacotron2 as autoregressive TTS architecture. For alignments, GradTTS, DVT, and VITS use hard alignments based on monotonic alignment search [18] and a duration predictor. Tacotron2 uses soft attention based on location-sensitive attention for alignments [15]. For the text encoder, GradTTS, DVT, and VITS use the transformer encoder, and Tacotron2 uses a CNN-based encoder. To clarify the effect of the encoder structure, we include DVT with a modified version of the textual encoder. We add three layers as a PreNet module to the original encoder: a 1D convolutional layer, a normalization layer followed by ReLU activation, and the layer applying dropout. We refer to this system as DVT2.

3. Experimental Evaluations

3.1. Experimental Conditions

Table 1 shows the noise condition of transcripts. To obtain noisy transcripts with ASR, we simulated noisy speech. We used audio from the DEMAND database [22] to simulate noise in the environment. We randomly selected the audio from DEMAND to be paired with the speech data from LJSpeech [23] and mixed the paired audio using different Signal-to-Noise Ratio (SNR) values to obtain artificially mixed noisy audio. To simulate noisy transcription data closer to the real situation, we set five SNR intervals: $[-10, 0)$, $[-15, 0)$, $[-20, 0)$, $[-20, -5)$, and

Transcript Type	SNR(dB) Interval	$CER\downarrow$ (%)	$CCorr\uparrow$ (%)	$WER\downarrow$ (%)	$WCorr\uparrow$ (%)	S.Err↓ (%)
Manual	-	0.0	100.0	0.0	100.0	0.0
WER-070	∞	3.3	97.6	7.0	94.3	59.6
WER-162	[-10, 0)	9.9	92.0	16.2	86.1	81.4
WER-242	[-15, 0)	15.7	86.9	24.2	79.2	86.8
WER-358	[-20, 0)	24.8	78.5	35.8	68.5	90.1
WER-437	[-20, -5)	30.9	73.1	43.7	61.4	94.2
WER-548	[-20, -10)	39.4	65.3	54.8	51.4	97.8

Table 1: *Noise Conditions of Transcriptions*

$[-20, 10)$. When the ambient and speech sounds were mixed, a random value was taken from the corresponding group of SNR intervals for mixing. This operation allows each group of speech to be mixed with noisy audio at different SNRs, resulting in mixed speech with varying noise levels.

We used ASR to obtain noisy transcripts from the simulated noisy speech. The noisy transcription results were in one-to-one correspondence with the SNR sets used in the mixing. We obtained five sets of noisy transcripts with 16.2, 24.2, 35.8, 43.7, and 54.8% word error rate (WER) results from the noisy speech with $[-10, 0)$, $[-15, 0)$, $[-20, 0)$, $[-20, -5)$, and $[-20, 10)$ SNR of mixing conditions, respectively. We denoted the obtained transcripts as WER-X where X indicated the WER results. We used a pre-trained model of Hybrid CTC/attention-based ASR [1] from ESPnet [2] for transcribing. This model achieved high recognition performance for the training set of original LJSpeech speech with 7.0% WER and 3.3% CER, and the transcribed results denoted as WER-070 into the experimental conditions. Finally, we incorporated manually annotated transcripts into the experimental conditions. In total, we investigated seven transcript conditions with different noise levels. In addition, characters and phonemes were used as two different types of transcripts and. Thus, there were 14 conditions of transcripts with varying noise levels and input representations. We split the simulated speech and noisy transcripts into train, validation, and test sets of 12,500, 100, and 250 samples.

We trained five TTS models to investigate the robustness to noisy transcription data : Tacotron2 [1], VITS [5], GradTTS [11], and DVT [8] and a variant of DVT. These methods used different textual encoder structures, probabilistic models for output speech, and alignments to relate input texts and output speech, which were expected to affect the robustness against noisy transcripts. Tacotron2 adopted a CNN-based text encoder, an autoregressive model of output speech, and a soft attention-based alignment method. VITS utilized a VAE-based textual encoder, flow-based probabilistic model of speech, and dynamic programming-based alignment method to align latent input and output representations in VAE. GradTTS consisted of a transformer-based text encoder, diffusion probabilistic model of speech, and dynamic programming-based alignment method to align the output of textual encoder with mel-spectrogram. DVT used a VAE-based textual encoder, diffusion probabilistic model of speech, and dynamic programming-based alignment method to align latent input and output representations in VAE. We also tested a modified version of the textual encoder for DVT. We added three layers as a PreNet module to the original encoder: a 1D convolutional layer, a normalization layer followed by ReLU activation , and a dropout layer. We referred to this system as DVT2.

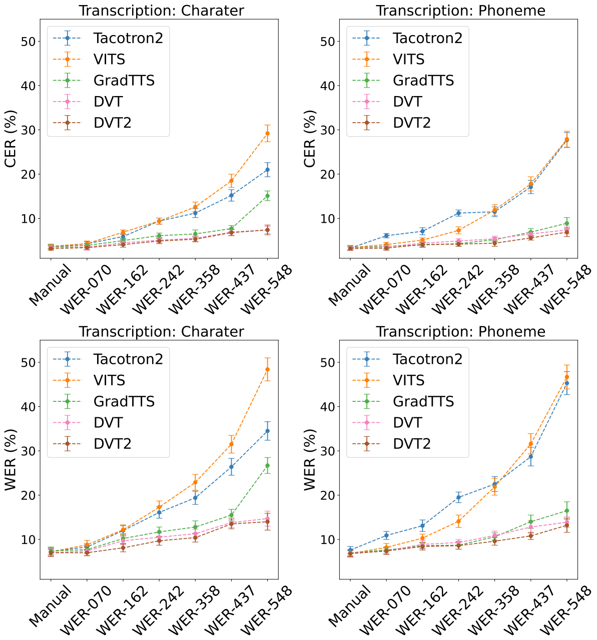

Figure 2: *Intelligibility Across Various Noise Levels*

We used clean speech to train the TTS systems. For Tacotron2 and GradTTS, we used mel-spectrogram as an acoustic feature and synthesized waveform from the mel-spectrogram with HiFiGAN vocoder [24]. For VITS, we used a linear spectrogram as an acoustic feature to train the acoustic encoder and HiFiGAN-based decoder in VAE. We synthesized waveforms in an end-to-end fashion with a jointly trained HiFiGAN-based decoder from latent representations. For DVT and DVT2, we used mel-spectrogram to separately train the HiFiGAN-based decoder to generate the waveform. After the training of the decoder, we trained the remaining model with the HiFiGAN-based decoder with frozen parameters. The five TTS methods were trained with the 14 different transcript conditions each. Therefore, we trained 70 TTS systems in total.

We used text samples from a test set of manually transcribed text in each model's synthesis and evaluated WER and CER of synthesized speech as objective intelligibility metrics [25]. We employed the same pre-trained ASR model that simulated noisy transcriptions. We checked the statistical significance of WER and CER differences between TTS methods with the same conditions with independent samples t-test with 95% confidence.

We conducted a listening test on naturalness to evaluate the subjective quality of synthetic speech. We included the five TTS methods under manual, WER-070, and WER-358 transcription conditions using phonemes in addition to natural samples. Each system contained 20 samples. We asked listeners to rate the audio samples on naturalness using a five-grade scale MOS with the correct transcript, prompting listeners to consider the effect of mispronunciation when scoring. We recruited 50 English-speaking listeners through Amazon Mechanical Turk. Each sample was evaluated by 20 listeners. We collected 6,400 evaluations in total. We checked the statistical significance of the MOS differences between TTS systems under the same transcription conditions.

[1] https://zenodo.org/records/3966501

[2] https://github.com/espnet/espnet

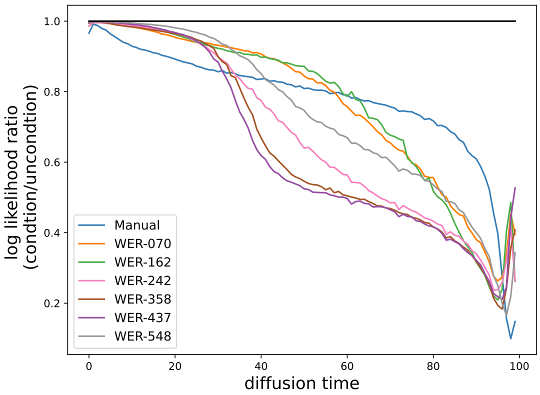

Figure 3: *Log-Likelihood Ratio of Phoneme-Conditioned to Unconditional DVT Model*

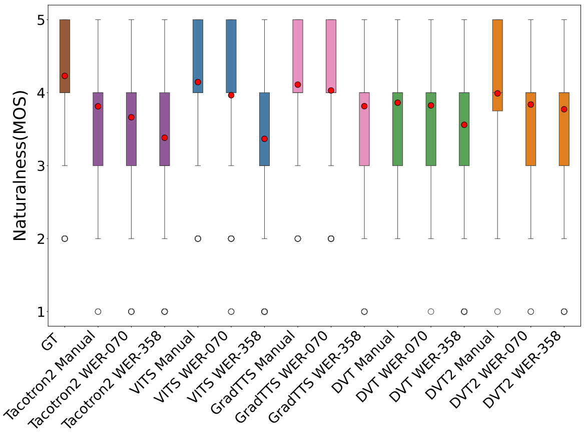

Figure 4: *MOS Scores for Naturalness*

3.2. Experimental Results

Figure 2 shows the result of objective evaluations on intelligibility. All methods showed similar performance under manual annotations, which were close to the recognition performance of ground truth. It indicated that all methods could render highly intelligible speech given normal transcripts. Under the manual annotation, there were little differences in intelligibility between character and phoneme representations. As the transcripts degraded, the TTS methods showed different results. Tacotron2 degraded its intelligibility greatly up to $45.3 \pm 2.6\%$ WER in phonemes as transcripts corrupted. It showed a greater degradation in phonemes than in characters. VITS showed intensive degradation as the transcript corrupted up to 46.7% WER in phonemes. Its intelligibility degraded similarly for both characters and phonemes. Those results indicated that autoregressive and flow-based TTS were not robust to transcript noise. GradTTS, on the other hand, showed a small degradation of intelligibility up to 16.5 ± 2.0 WER in phonemes as transcripts were corrupted. GradTTS also showed more significant degradation of intelligibility in characters than in phonemes. DVT showed similar trends to GradTTS with a smaller degree of degradation up to 13.9 ± 1.2 WER in phonemes. It showed an equal degradation between characters and phonemes, consistent with the reported robustness of ambiguous orthography [8]. DVT2 mitigated the intelligibility degradation slightly compared to DVT. It indicated that the network structure of the text encoder contributed to the robustness. These results supported that diffusion-based TTS were generally robust to transcript corruption.

Figure 3 shows histories of the likelihood ratio between the conditional and unconditional generation of DVT during inference clarifying the model's dependence on transcripts under noise conditions. The lower the ratio, the lower the likelihood of conditional generation was, indicating inferred latent speech was not close to ground truth. Note that the zero diffusion time was the final time step of inference. Given normal transcriptions, the model's likelihood was low initially, but it immediately improved after several steps and gradually approached 1.0 afterward. It indicated that the model generated the latent speech close to ground truth within early time steps under normal conditions and consumed most diffusion time for little improvement. In contrast, the models trained with noisy transcripts stayed low likelihood up to half diffusion time and improved likelihood abruptly afterward to approach 1.0. It indicated that the models took a long time to generate latent speech close to ground truth under noisy transcript conditions. These results suggested that training and inference with a long diffusion time were important for the robustness of transcription noise.

Figure 4 shows the results of the listening test. All methods showed high naturalness, from 3.8 to 4.1 in MOS, under manual annotations. Tacotron2 and VITS showed high degradation of naturalness as transcript corrupted: VITS degraded MOS from 4.15 ± 0.07 to 3.37 ± 0.16, and Tacotron2 degraded MOS from 3.81 ± 0.06 to 3.38 ± 0.17 at WER-358. On the other hand, diffusion-based TTS such as GradTTS, DVT, and DVT2 showed only slight degradation of naturalness as transcript corrupted: GradTTS dropped MOS from 4.11 ± 0.10 to 3.82 ± 0.12, DVT dropped MOS from 3.87 ± 0.10 to 3.56 ± 0.17 at WER-358. Compared to DVT, DVT2 could mitigate the degradation more from 3.99 ± 0.09 to 3.77 ± 0.15 at WER-358. These results supported the robustness of diffusion-based TTS in human perception as well.

4. Conclusion

This paper investigated the robustness of text-to-speech (TTS) models with different architectures in the face of noisy transcriptions with various levels of corruption. We obtained noisy transcripts using automatic speech recognition on simulated noisy speech. We evaluated autoregressive, flow, and diffusion-based models trained on noisy transcriptions under different conditions regarding objective intelligibility and subjective naturalness and compared them with models trained on manual transcriptions. Our experimental results showed that the diffusion-based TTS model was extremely robust to severely corrupted transcription data, mitigating about 30% of the word error rate compared to autoregressive and flow-based models. We revealed that iterative inference with a long diffusion time contributed to the robustness of diffusion-based TTS based on likelihood analysis. Among diffusion-based TTS methods, DVT2 performed best in terms of high intelligibility under noisy characters and phonemes and high naturalness, thanks to the encoder structure and variational autoencoder.

In future work, we will analyze the robustness of TTS under noisy transcripts during inference in addition to training.

5. Acknowledgements

This work was supported in part by project JPNP20006 commissioned by NEDO, and by JSPS KAKENHI Grant Number 21H05054. Additionally, this work received financial support from JST SPRING, Grant Number JPMJSP2125 and we would like to thank the "THERS Make New Standards Program for the Next Generation Researchers."

6. References

[1] 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.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[4] 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," in *International Conference on Learning Representations*, 2021.

[5] 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.

[6] D. Ma, Z. Su, Y. Zhang, E. Huang, M. Li, Q. Lyu, and F. Ye, "Mhtts: Fast multi-head text-to-speech for spontaneous speech with imperfect transcription," in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2022, pp. 239–244.

[7] J. Fong, P. O. Gallegos, Z. Hodari, and S. King, "Investigating the Robustness of Sequence-to-Sequence Text-to-Speech Models to Imperfectly-Transcribed Training Data," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 1546–1550.

[8] Y. Yasuda and T. Toda, "Text-to-speech synthesis based on latent variable conversion using diffusion probabilistic model and variational autoencoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] H. Kim, S. Kim, and S. Yoon, "Guided-tts: A diffusion model for text-to-speech via classifier guidance," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 11 119–11 133.

[10] J. Tae, H. Kim, and T. Kim, "Editts: Score-based editing for controllable text-to-speech," in *INTERSPEECH*. ISCA, 2022, pp. 421–425.

[11] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8599–8608.

[12] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[14] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds., 2015, pp. 1412–1421.

[15] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 577–585.

[16] Y. Yasuda, X. Wang, and J. Yamagishi, "Initial investigation of encoder-decoder end-to-end TTS using marginalization of monotonic hard alignments," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 211–216.

[17] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "Alignts: Efficient feed-forward text-to-speech system without explicit alignment," in *ICASSP*. IEEE, 2020, pp. 6714–6718.

[18] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *NeurIPS*, 2020.

[19] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[20] Y. Kim, Y. Jernite, D. A. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 2741–2749.

[21] Y. Yasuda, X. Wang, and J. Yamagishi, "Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis," *Computer Speech & Language*, vol. 67, p. 101183, 2021.

[22] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND)*, Montreal, Canada, 2013, pp. 035 081–035 081.

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

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[25] J. Taylor and K. Richmond, "Confidence intervals for asr-based tts evaluation." in *Interspeech*, 2021, pp. 2791–2795.