Expressive, Variable, and Controllable Duration Modelling in TTS

Ammar Abbas, Thomas Merritt, Alexis Moinet, Sri Karlapati, Ewa Muszynska, Simon Slangen†, Elia Gatti, Thomas Drugman

Alexa AI, Amazon

syeabbs@amazon.co.uk

Abstract

Duration modelling has become an important research problem once more with the rise of non-attention neural text-to-speech systems. The current approaches largely fall back to relying on previous statistical parametric speech synthesis technology for duration prediction, which poorly models the expressiveness and variability in speech. In this paper, we propose two alternate approaches to improve duration modelling. First, we propose a duration model conditioned on phrasing that improves the predicted durations and provides better modelling of pauses. We show that the duration model conditioned on phrasing improves the naturalness of speech over our baseline duration model. Second, we also propose a multi-speaker duration model called Cauliflow, that uses normalising flows to predict durations that better match the complex target duration distribution. Cauliflow performs on par with our other proposed duration model in terms of naturalness, whilst providing variable durations for the same prompt and variable levels of expressiveness. Lastly, we propose to condition Cauliflow on parameters that provide an intuitive control of the pacing and pausing in the synthesised speech in a novel way.

Index Terms: neural text-to-speech, normalising flows, expressive TTS, duration modelling

1. Introduction

Text-to-speech (TTS) systems have seen many paradigm shifts in recent years. Neural TTS systems achieved state-of-the-art results in speech synthesis with different architectures [1–5]. However, they lack robustness to out-of-distribution text, thus causing issues such as mumbling, skipping, or repetition in speech. Therefore, various new methods have been proposed that are more robust to unseen text utterances [6, 7]. They employ duration models to find alignment between input phonemes and target acoustic features. Many recent methods [8–10] have been based on this architecture, and have gradually improved the acoustic models. At the same time, the duration models have largely not been updated since the statistical parametric speech synthesis approaches. These duration models are deterministic, and generally trained via distance-based losses like L1/L2 loss without any additional conditioning. This results in close to mean duration prediction in speech with limited expressiveness. These uni-modal predicted durations also lead to insufficient pauses in longer texts, adversely affecting the listener's understanding of the text [11]. Klimkov *et al.* [12] and Liu *et al.* [13] share a phrasing model that is introduced to address this shortcoming, however, they are presented for an attention based TTS system.

Variability in the synthesised speech is another important factor in order to improve the perceived naturalness of the TTS systems for long and repetitive texts such as question-answering prompts. This is achievable using some of the recent generative modelling techniques such as normalising flows also known as flows. Flows [14] provide a tractable solution for maximising exact log-likelihood which allows predicting distributions that are true to the target distribution. Furthermore, they allow variations in the output by controlling the prior. Flows have recently attracted attention in the TTS domain for modelling diverse acoustic features [15–17]. However, the duration models used are still deterministic, and trained with a distance-based loss, thus suffering from the aforementioned limitations. Kim *et al.* [18] describe a similar approach to ours by using a stochastic duration model, however, they obtain the target durations for their duration model via Monotonic Alignment Search (MAS). MAS can suffer from attention instabilities when extracting the target durations during training [15]. Additionally, their duration model is conditioned only on phoneme embeddings, and has a significantly different architecture.

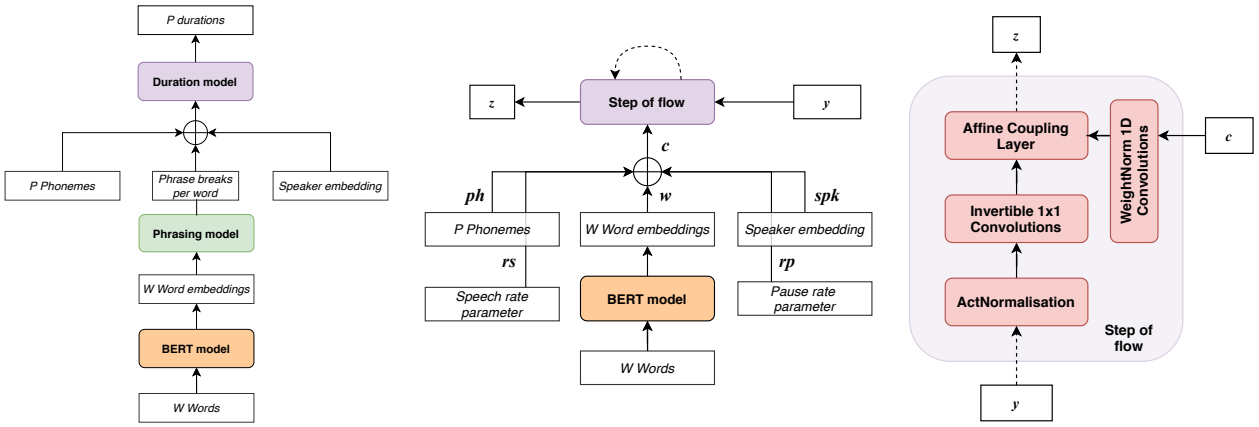
In this paper, we present two alternate duration modelling approaches: (a) duration model conditioned on phrasing; (b) Cauliflow. Our main contributions are: (i) We show that duration model conditioned on phrasing is statistically significantly preferred to the baseline duration model; (ii) We explore the performance of the duration model conditioned on phrasing, and show that it improves the predicted durations, pause rate, speech rate, and $\mathcal{F}_0.25$ score over the baseline duration model; (iii) We propose a flow-based multi-speaker stochastic duration model called Cauliflow (Conditional Alignment Using normaLIsing FLOWs) that is conditioned on BERT word embeddings to predict semantic and syntax informed durations. Cauliflow performs on par with duration model conditioned with phrasing in terms of naturalness; and finally (iv) We show that Cauliflow predicts more expressive durations, and can produce variable durations for the same text. Additionally, we show that the proposed novel parameters in Cauliflow control the pace and pausing in the speech in an intuitive way.

2. Baseline

2.1. Acoustic Model

We train an acoustic model that is a multi-speaker version of the baseline introduced in [19]. The acoustic model is trained separately from the duration model. This allows us to compare different duration models given the same acoustic model. First, we obtain P phoneme embeddings $\boldsymbol{ph} = [ph_0, ph_1, \cdots, ph_{P-1}]$ from the encoder given input text, where P is the number of phonemes. The phoneme embeddings $\boldsymbol{ph}$ are then upsampled according to the durations $\boldsymbol{d} = [d_0, d_1, \cdots, d_{P-1}]$, such that $\sum_{i=0}^{P-1} d_i = T$, where T is the total number of mel-spectrogram frames. The upsampled phoneme embeddings are passed through the decoder to obtain the mel-spectrogram. We use a universal neural vocoder [20] to obtain speech signal from the generated mel-spectrogram.

† The work was done while author was at Amazon.



(a) *Duration model conditioned on phrasing* (b) *Cauliflow model* (c) *Inside view of a step of flow*

Figure 1: *Comparison of the architecture of Cauliflow and duration model conditioned on phrasing. (b) and (c) shown at training time.*

2.2. Duration model

We train a duration model that is optimised with the L2 loss to predict normalised durations $\boldsymbol{d}$, a common approach in many duration based architectures [9, 15, 16]. All duration models in this paper are trained using the same oracle durations, obtained from Montreal Forced Alignment [21] for a fair comparison.

3. Duration model conditioned on phrasing

The baseline duration model is trained with the L2 loss that has some limitations in practice. The local minimum converges to the mean of the data [22] which can result in monotonous durations if any additional conditioning is not provided, thus producing flatter or less expressive speech. This particularly affects the prediction of the occurrence and duration of pauses. Indeed, in our data, most pauses occur at punctuation marks, but more than 20% can be found between words not separated by a punctuation mark, i.e. a word boundary. However, only around 2.5% of word boundaries correspond to a pause. Thus due to the averaging nature of the L2 loss, our baseline model predicts that the duration at any word boundary approaches 0 (i.e. no pause), unless there is a punctuation mark.

To better model the distribution of pauses in the data, we train a duration model that is conditioned on output from a phrasing model. The phrasing model is a classification model trained with Cross Entropy loss against target binary phrasing labels that are obtained from forced alignment. The phrasing model architecture can be seen in Figure 1a. The input text is passed through a fine-tuned BERT-base model [23]. The BERT model predicts sub-word level embeddings from the input text, which are then averaged to compute word-level BERT embeddings passed to the phrasing model.

The output of the phrasing model is the probability of a pause after each word that is converted into a binary value, i.e., occurrence of a pause, before being passed to the duration model. We choose a threshold that optimises the $\mathcal{F}_{0.25}$ score [12] for predicted pauses in order to obtain higher precision for a better listening experience. During training, the duration model is conditioned on oracle phrasing labels. During inference, it is conditioned on the output of the phrasing model.

4. Cauliflow

Cauliflow is a normalising flow based duration model that is conditioned on phoneme, BERT and speaker embeddings. Additional conditioning on speech and pause rate allows us to further control the predicted duration distribution.

Normalising flows allow optimising for the exact log-likelihood of the target distribution. Let $\boldsymbol{z}$ be the prior Gaussian distribution with the probability density function $P(\boldsymbol{z})$, and let $P(\boldsymbol{y}|\boldsymbol{c})$ be the target duration $\boldsymbol{y}$ distribution given conditional features $\boldsymbol{c}$. Normalising flows are composed of series of invertible functions $\boldsymbol{f}(\boldsymbol{z}|\boldsymbol{c}) = f_0 \circ f_1 \circ f_2 \cdots f_k$ that transform the Gaussian distribution $\boldsymbol{z}$ into the target distribution $\boldsymbol{y}$:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{z}|\boldsymbol{c}) = f_0 \circ f_1 \circ f_2 \cdots f_k(\boldsymbol{z}|\boldsymbol{c}), \tag{1}$$

then by a change of variables, we can compute the log-likelihood of the target distribution $\boldsymbol{y}$ as:

$$\log P_Y(\boldsymbol{y}|\boldsymbol{c}) = \log P_Z(\boldsymbol{z}) + \sum_{i=1}^{K} \log |\det(\boldsymbol{J}(f_i^{-1}(\boldsymbol{x}|\boldsymbol{c})))|, \tag{2}$$

where $\log|(\det(\boldsymbol{J}))|$ represents the log-value of the determinant of the Jacobian matrix $\boldsymbol{J}$.

Cauliflow is specifically based on the architecture of Flow-TTS [16]. An overview of Cauliflow model can be seen in Figure 1b. The conditions $\boldsymbol{c}$ are added in the affine coupling layers [24] in the flow steps as shown in Figure 1c, according to equation 1. We now detail the steps required to compute different parameters in $\boldsymbol{c}$.

4.1. Phoneme embeddings

The phoneme embeddings $\boldsymbol{ph}$ contain information about the content to be synthesised. They are obtained by passing the phonemes $\boldsymbol{p}$ through an encoder with an architecture similar to the encoder used in Tacotron 2 [2].

4.2. BERT embeddings

The semantic and syntactical information in a sentence plays an important role in the prosody of the synthesised speech [25, 26]. BERT embeddings have been shown to contain information about the semantic and syntactical structure of a sentence [27]. To leverage that information to guide duration prediction, we condition Cauliflow on embeddings obtained from BERT. First, we obtain word-level BERT embeddings using the process described in Section 3. The word-level BERT embeddings are then upsampled according to the number of phonemes in each word to compute phoneme-level BERT embeddings $\boldsymbol{w}$.

Speaker	System	JSD		Punctuations			Word boundaries			Pause rate	Speech rate
		Pause	Non-pause	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_{0.25}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_{0.25}$		
A	**Target**	-	-	-	-	-	-	-	-	5.86	3.07
	Dur	0.36	0.09	92.27	92.78	92.30	72.28	12.71	56.65	6.98	3.32
	Dur+P	0.25	0.07	91.07	98.66	91.48	65.63	64.32	65.55	5.66	3.10
	Cauliflow	0.19	0.03	95.01	83.66	94.25	66.58	46.82	64.97	6.30	3.11
B	**Target**	-	-	-	-	-	-	-	-	7.65	3.59
	Dur	0.56	0.09	82.29	48.19	79.00	0.0	0.0	0.0	10.53	3.94
	Dur+P	0.26	0.07	76.93	95.52	77.82	42.80	48.13	43.08	6.85	3.48
	Cauliflow	0.15	0.04	82.76	70.59	81.93	33.21	37.76	33.45	7.81	3.46

Table 1: *Comparison of objective metrics between different durations models. Pause and speech rate are better when closer to Target.*

4.3. Speaker embeddings

The speaker embeddings $\boldsymbol{spk}$ are obtained from a speaker verification system that is trained using a network based on Generalised End-to-End loss [28]. A speaker embedding is a 192-dimensional vector computed for each utterance during training. An utterance can be composed of one or more sentences. During inference, we use a mean speaker embedding that is computed given all the training utterances for a particular speaker.

4.4. Speech and pause rate

A parameter $\boldsymbol{rs}$ is introduced to control the pace of the speech. Let μ_{rs} be the average number of words per second in the training data. Then the model is conditioned on $\boldsymbol{rs}$, calculated for each utterance as:

$$\boldsymbol{rs} = \frac{W}{D} - \mu_{rs}, \tag{3}$$

where W is the number of words, and D is the duration of the utterance in seconds. The model learns the relation between $\boldsymbol{rs}$ and the speech rate to be controlled in the synthesised speech, i.e., the scalar value 0 represents the average number of words per second seen in the data, while positive or negative values of $\boldsymbol{rs}$ represent a faster or slower speech respectively. A similar approach is followed for the parameter $\boldsymbol{rp}$ controlling the pause rate which is calculated as:

$$\boldsymbol{rp} = \frac{W}{S} - \mu_{rp}, \tag{4}$$

where S is the number of pauses in the utterance, and μ_{rp} is the average number of words per pause in an utterance in the training data.

Given these different parameters, Equation 1 for normalising flows can be expanded to:

$$\boldsymbol{y} = f(\boldsymbol{z}|\boldsymbol{c}) = f(\boldsymbol{z}|\boldsymbol{ph}, \boldsymbol{w}, \boldsymbol{spk}, \boldsymbol{rp}, \boldsymbol{rs}) \tag{5}$$

The target durations $\boldsymbol{y}$ are not normalised in Cauliflow. We experimented with different normalisation schemes such as log-normalisation and z-score, but found that non-normalised target durations give the best results. This could be due to possible outliers which are discussed in Section 5.4.

5. Experimental Validation

We evaluate three different duration models: a) the baseline from Section 2.2, referred as "Dur"; b) duration model conditioned on phrasing from Section 3, referred as "Dur+P"; c) Cauliflow from Section 4. The ground truth durations are referred as "Target".

5.1. Data

Experiments are conducted on an internal English voice dataset consisting of 1 male and 3 female speakers. The models are trained on all 4 voices comprising 50 hours of total training data, while the test set consists of around 3 hours of data for each speaker. We report our results on a subset of two speakers chosen randomly, referred from hereon as speaker A and speaker B. The sampling rate of the recorded audio is 24 kHz. We extract 80 band mel-spectrograms from the audio with a frame shift of 12.5ms.

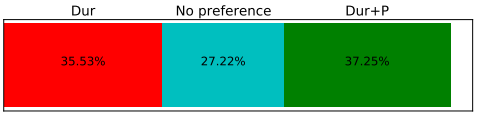


Figure 2: *Preference test result shows that Dur+P is statistically signficantly better than Dur on speaker A (p-$val < 0.05$).*

	Dur+P	Cauliflow	CopySynth
Speaker A	73.58±1.10	73.53±1.10	76.03±1.02
Speaker B	73.62±1.13	73.22±1.14	74.21±1.12

Table 2: *Comparison of MUSHRA scores between different systems along with their 95% confidence intervals.*

5.2. Comparing Dur+P against Dur

We conducted a preference test between Dur and Dur+P. The evaluation consisted of 50 utterances randomly selected from speaker A test set, and were rated by 120 native English listeners. The results are shown in Figure 2. We use a binomial significance test to measure the statistical significance, and find that Dur+P is statistically significantly better than Dur (p-val$<$0.05).

To gain more insights, we also compare the durations predicted by both models on the whole test set as shown in Table 1, using different objective metrics such as speech and pause rate, precision $\mathcal{P}$, recall $\mathcal{R}$, and $\mathcal{F}_{0.25}$ score [12] over pauses, and Jensen-Shannon divergence (JSD). We choose JSD as a metric because it measures the closeness between two distributions in a symmetric way. The comparison shows that Dur+P has a speech and pause rate that is closer to the speaker's statistics. The Dur model does not predict a pause on any word boundary for speaker B, while Dur+P significantly improves the $F_{0.25}$ for both speakers. We also note that Dur+P has lower JSD on predicted duration distribution for both speakers on pause and non-pause tokens. We believe that the model is also performing better on non-pause tokens because information about pause occurrence can affect the duration of adjacent phonemes.

5.3. Comparing Cauliflow against Dur+P

In this section, we evaluate both of our proposed duration models, Cauliflow and Dur+P. We run a MUSHRA [29] evaluation consisting of Cauliflow, Dur+P, and copy-synthesis of the original recordings (CopySynth) to quantify improvements in naturalness of the systems. Copy-synthesis is used to focus only on improvements from the duration model and its effect on the acoustic model. The evaluations are run for speaker A and B

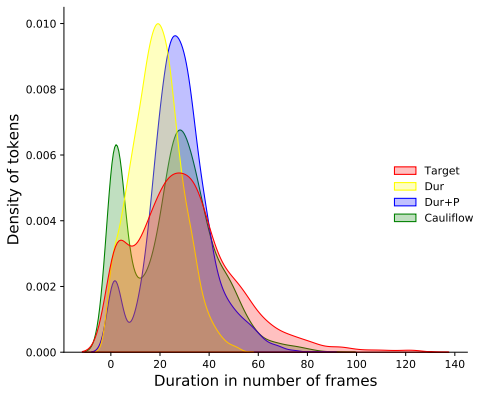


Figure 3: *Comparison of predicted durations on punctuation marks by different duration models.*

System	JSD		99th Percentile L1 Error (frames)
	Pauses	Non-pauses	
Cauliflow ($\mathcal{T} = 0.3$)	0.21	0.05	28
Cauliflow ($\mathcal{T} = 0.5$)	0.20	0.04	29
Cauliflow ($\mathcal{T} = 0.7$)	0.19	0.03	31
Cauliflow ($\mathcal{T} = 1.0$)	0.19	0.03	34

Table 3: *Comparison of duration metrics for various $\mathcal{T}$.*

consisting of 50 utterances each rated by a total of 240 native English speakers. The results of the evaluations can be seen in Table 2. We run a pairwise two-sided Wilcoxon signed-rank test, corrected for multiple comparisons to measure statistical significance between the systems. We found that Cauliflow and Dur+P are on par in terms of naturalness on both speakers (p-val > 0.05). Both the systems obtain MUSHRA scores that are very close to the MUSHRA score for CopySynth on speaker B, with no statistically significant difference between Dur+P and CopySynth.

To understand more about how the predicted durations differ between both the models, we show their objective metrics in Table 1. We note that Cauliflow generally has a better precision, while Dur+P has a better recall on predicted pauses, thus providing use cases for different situations. The predicted durations from Cauliflow have lower JSD with the target durations compared to Dur+P on both pauses and non-pause phonemes for both speakers. This can also be seen by the duration distribution of these methods over the punctuation tokens in Figure 3. The Dur+P model has a higher mode whereas Cauliflow has a better bi-modal distribution of durations that denotes pauses vs non-pauses. The tails of the Cauliflow distribution also show better distribution coverage but can sometimes result in outlier predictions. This shows that Cauliflow produces a more natural duration distribution but it can suffer from outlier predictions. We further explore the latter in detail in the next section.

5.4. Controlling duration variability

We evaluate the effect of temperature $\mathcal{T}$ in the prior Gaussian distribution $\mathcal{N}(0, \sigma.\mathcal{T})$ on the quality of predicted durations from Cauliflow. We use a prior Gaussian distribution with parameters $\mathcal{N}(0, 1)$ during training. However, we note that with $\mathcal{T} = 1$, the model predicts durations that can result in unstable speech synthesis. It is partly attributed to the outliers identified by computing different percentile errors. The 99th percentile of the errors in predicted durations vs target durations is shown in Table 3. With a lower value of $\mathcal{T}$, we note that there are fewer outliers, however, the model starts to predict worse duration distributions as shown by the higher JSD in the table. We find that $\mathcal{T} = 0.7$ is the right trade-off between a better predicted duration distribution and minimizing the outlier predictions. In the the future, we will investigate if there is a way to minimize outlier predictions while still using $\mathcal{T} = 1$.

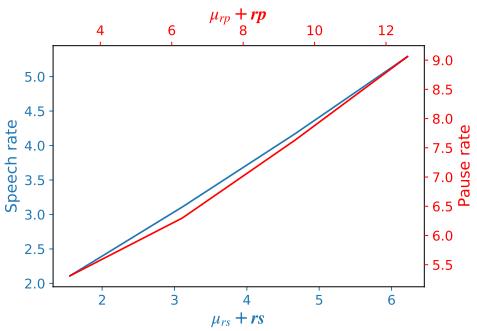


Figure 4: *Correlation of parameters for speech and pause rate versus their measured impact on speech.*

"Are you reading <P1> or just wondering <P3> that whatever happens here <P4> around this time <P2> will stay here forever?"

Table 4: *Insertion of more pauses in a textual prompt in the order <P1>, <P2>, <P3>, <P4>.*

5.5. Controlling speech and pause rate via Cauliflow

We vary the values of $\boldsymbol{rp}$ and $\boldsymbol{rs}$ to measure their effect on the pace and pausing of the synthesised speech on 3 hours of test set of speaker A. The resultant effect can be seen in Figure 4. We find that that there is a direct correlation between control parameters versus the speech and pause rate. Furthermore, the change in pacing and pausing is close to the intended value of the parameters while being a bit conservative, i.e., the decrease or increase in the speech or pause rate is generally less than the intended value. In the case of pause rate, we note that this is generally because the placement of pauses follows the distribution seen in the data, and thus doesn't put pauses at inappropriate locations in the text. An example of varying pause rates for a sentence can be seen in Figure 4. We vary $\boldsymbol{rp}$ decreasing it to as low as around one word per pause, but note that once all semantically appropriate pauses are placed in the text, the model doesn't place any more pauses. Furthermore, new pauses are placed in the decreasing order of their suitability to text ,i.e., pause <P1>is placed when using the default $\boldsymbol{rp}$, <P2>is placed when decreasing it by around 1 word per pause, and so on. Therefore, we believe that these parameters can be safely used in practice as they generally avoid pause insertion that could result in a bad listening experience.

6. Conclusion

In this paper, we proposed a modification to duration models that incorporates a phrasing model, and showed that it improves the predicted durations and produces more natural speech over the baseline. We also proposed a normalising flow based multi-speaker duration model called Cauliflow that is conditioned on BERT embeddings for more appropriate duration prediction. We showed that Cauliflow performs on par with the duration model conditioned on phrasing, while providing variable durations, and providing an intuitive way to control the speech and pause rate of the synthesised speech.

7. References

[1] 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," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

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

[3] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning." in *ICLR (Poster)*, 2018.

[4] 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," *9th ISCA Speech Synthesis Workshop*, 2016.

[5] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[6] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "DurIAN: Duration informed attention network for multimodal synthesis," in *Interspeech*, 2020.

[7] 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.

[8] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "AlignTTS: Efficient feed-forward text-to-speech system without explicit alignment," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 6714–6718.

[9] 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*, 2020.

[10] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Skerry-Ryan, and Y. Wu, "Parallel Tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling," *arXiv preprint arXiv:2103.14574*, 2021.

[11] S. R. Rochester, "The significance of pauses in spontaneous speech," *Journal of Psycholinguistic Research*, vol. 2, no. 1, pp. 51–81, 1973.

[12] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, "Phrase Break Prediction for Long-Form Reading TTS: Exploiting Text Structure Information." in *Interspeech*, 2017, pp. 1064–1068.

[13] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "Modeling prosodic phrasing with multi-task learning in tacotron-based TTS," *IEEE Signal Processing Letters*, vol. 27, pp. 1470–1474, 2020.

[14] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[15] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[16] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A non-autoregressive network for text to speech based on flow," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7209–7213.

[17] R. Valle, K. J. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis," in *International Conference on Learning Representations*, 2020.

[18] 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.

[19] S. Karlapati, A. Abbas, Z. Hodari, A. Moinet, A. Joly, P. Karanasou, and T. Drugman, "Prosodic representation learning and contextual sampling for neural text-to-speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6573–6577.

[20] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal neural vocoding with parallel wavenet," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6044–6048.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[22] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[23] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[24] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," in *ICLR (Poster)*, 2017.

[25] R. Bennett and E. Elfner, "The syntax–prosody interface," *Annual Review of Linguistics*, vol. 5, pp. 151–171, 2019.

[26] L. C. Nygaard, D. S. Herold, and L. L. Namy, "The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning," *Cognitive science*, vol. 33, no. 1, pp. 127–146, 2009.

[27] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020.

[28] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[29] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.