Controllable Neural Prosody Synthesis

Max Morrison[1,*], *Zeyu Jin*[2], *Justin Salamon*[2], *Nicholas J. Bryan*[2], *Gautham J. Mysore*[2]

[1]Northwestern University, Evanston, IL, USA
[2]Adobe Research, San Francisco, CA, USA

maxrmorrison@gmail.com, zejin@adobe.com

Abstract

Speech synthesis has recently seen significant improvements in fidelity, driven by the advent of neural vocoders and neural prosody generators. However, these systems lack intuitive user controls over prosody, making them unable to rectify prosody errors (e.g., misplaced emphases and contextually inappropriate emotions) or generate prosodies with diverse speaker excitement levels and emotions. We address these limitations with a user-controllable, context-aware neural prosody generator. Given a real or synthesized speech recording, our model allows a user to input prosody constraints for certain time frames and generates the remaining time frames from input text and contextual prosody. We also propose a pitch-shifting neural vocoder to modify input speech to match the synthesized prosody. Through objective and subjective evaluations we show that we can successfully incorporate user control into our prosody generation model without sacrificing the overall naturalness of the synthesized speech.

Index Terms: prosody generation, speech editing, speech synthesis, text to speech, voice modification, vocoder

1. Introduction

Text is a strong indicator of prosodic patterns [1], but not a determinant. For the same text, prosody varies with speaker intention [2], which imposes challenges for modern text-to-speech models [3, 4, 5] in the form of misplaced emphases and degraded naturalness. Manually attempting such corrections using an audio editor, as is done in podcast and video dialogue editing, requires expertise in speech manipulation and significant time and effort. In this paper, we aim to address these issues by proposing an intuitive and less error-prone process, consisting of three steps: (1) a user provides constraints on the prosody (e.g., by drawing part of an F0 contour), (2) a neural prosody generator predicts an F0 contour for the whole utterance while matching the user's constraints, and (3) a neural vocoder synthesizes a high-fidelity speech recording that exhibits the generated prosody.

Early approaches for F0 synthesis use techniques such as decision trees [6], unit selection [7], and hidden Markov models (HMMs) [8]. More recently, deep learning methods such as variational autoencoders (VAEs) [9], deep autoregressive (DAR) neural networks [10], and vector-quantized VAEs (VQ-VAEs) [11] were shown to be effective at generating F0 contours of speech from text features. Hodari et. al. [9] show that VAEs produce F0 contours that cluster around placing emphasis on the same words despite repeated sampling. This indicates that the VAE is not capturing the multimodal nature of English prosody associated with contrastive emphases. For example, the sentence "the dog is black" communicates a different intention when one of "dog", "is", or "black" is emphasized. The DAR model proposed in [12] has previously shown promise in modeling the multimodality of English prosody [10] but does not allow user control over F0 generation. While our work focuses on user control of F0 generation, additional prosodic control can be achieved by first generating a speech waveform with the desired phoneme durations (e.g., with [5]) and then using our method to achieve the desired F0.

Once we generate an F0 contour, we synthesize speech using a compatible vocoder. Existing vocoders allow either perceptually high-quality synthesis [3, 13, 14] or a high degree of control over prosody [15]. Recently, significant effort has gone into disentangling the latent spaces of high-quality neural vocoders to recover explicit prosodic control [16, 17, 18, 19]. One such model, Quasi-Periodic WaveNet, allows frame-wise F0 control via an explicit F0 contour but produces lower naturalness than DSP-based vocoders, especially when pitch is shifted upward [20]. In contrast, we propose a pitch-shifting neural vocoder that achieves comparable or superior performance as DSP-based methods while factorizing prosody control parameters in the input space using a simple, jointly-trained bottleneck.

Our key contributions are: (1) a novel method for F0 generation that permits intuitive user controls, (2) a pitch-shifting neural vocoder with explicit F0 conditioning, and (3) a new subjective evaluation method for measuring the naturalness of prosody. Through our perceptual evaluation, we show that user control of prosody can be obtained without degrading prosody naturalness, and our pitch-shifting neural vocoder performs comparably with existing DSP-based methods while outperforming prior neural pitch-shifting methods.

2. Controllable F0 generation

2.1. Deep autoregressive (DAR) neural network

For its effectiveness and simplicity, we use DAR as our baseline model for F0 generation (model $\texttt{QF}_{\texttt{FT}}$ [12]). DAR feeds input text features through two fully-connected layers with ReLU activation followed by two RNNs, one bidirectional and one unidirectional, followed by a fully-connected layer. The outputs of the last fully-connected layer are the logits of a categorical distribution of quantized F0 values. One F0 value is sampled per time frame and the resulting observation (a one-hot-encoded F0 value) is concatenated to the input of the unidirectional RNN at the next frame (i.e., the unidirectional RNN is autoregressive). To prevent the unidirectional RNN from ignoring the current input features and focusing on its hidden state (i.e., exposure bias), *data dropout* is used, whereby autoregressive inputs to the unidirectional RNN are set to zero with probability p (we use $p = 0.5$, as in [12]). The original DAR uses a hierarchical softmax loss to improve binary classification of voiced/unvoiced (V/UV) frames. However, the ground truth V/UV sequence can also be derived directly from phonemes when synthesizing

*This work was carried out during an internship at Adobe Research.

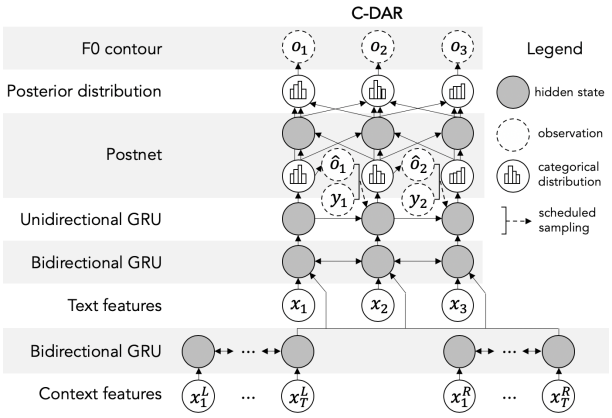

Figure 1: *Our proposed C-DAR architecture for controllable F0 generation.* x_t, y_t, o_t, *and* $\hat{o}_t$ *are the input features, ground truth F0, predicted F0, and predicted F0 before postnet, respectively, and* $t = 1, \ldots, T$ *is the current frame.* x_t^L *and* x_t^R *are the input features from the preceding and following speech. The fully-connected layers between the text features and the bidirectional GRU as well as three layers of the postnet are omitted.*

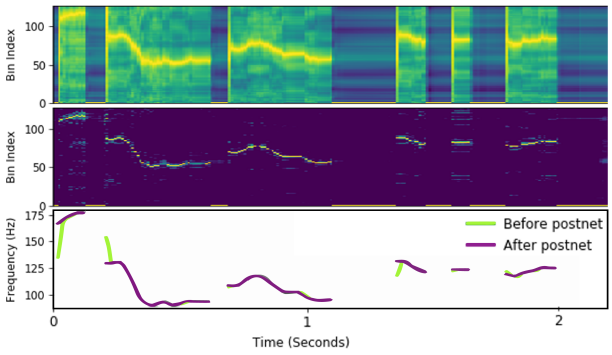

Figure 2: *Effects of postnet on F0 generation.* ***(Top)*** *Log posterior distribution at each frame before postnet.* ***(Middle)*** *Log posterior distribution after postnet.* ***(Bottom)*** *Argmax of each distribution converted to frequency. Posteriors are per-frame normalized to have a maximum of 0 and clipped below -40.*

speech, or from a preexisting F0 contour when editing speech. We use V/UV sequences from existing F0 contours and therefore use cross-entropy loss instead of hierarchical softmax.

DAR has been shown to be effective at modeling English prosody, but does not permit user control and lacks the context-awareness necessary for speech editing tasks. We address these limitations in our proposed F0 generation model, Controllable DAR (C-DAR), shown in Figure 1.

2.2. Controllable DAR

A significant advantage of working with an F0 contour, as opposed to jointly predicting all prosodic features, is that users may explicitly create, modify, and constrain the F0 contour to realize a creative goal. We propose three techniques designed to facilitate control of a DAR-based model for F0 generation: (1) if available, the preceding and following speech content is summarized and used to condition the model, (2) random segments of the ground truth F0 are provided to the model during training, and (3) the model predicts F0 values in reverse order.

The preceding and following speech content provides useful indicators for placing emphases [2], capturing the speaker's current prosodic style [21], and determining F0 values near boundary points. This context-awareness is essential in speech editing tasks, where prosody edits must sound natural relative to the surrounding speech. We incorporate context-awareness by summarizing the preceding and following content each with an untied, two-layer bidirectional GRU with hidden size 128. We use the same input features for the preceding and following content (see Section 2.3) with the addition of one-hot-encoded F0 values. The result is concatenated with the text features at the input of the model at each time frame.

A potentially useful user interaction for controllable F0 generation is to explicitly specify some segments of the desired contour (e.g., by placing and moving anchor points or drawing) and have a generative model infer the remaining F0 values. This permits iterative refinement, in which a user generates an F0 contour using our model, selects regions they want to keep, regenerates only unselected regions, and repeats until satisfied. Explicitly specifying a higher pitch over a word also allows users to quickly create emphases[1]. This is useful as emphases in English are challenging to predict [22] as they can arise semantically [23] or simply due to speaker excitement [21]. We implement this technique by providing ground-truth F0 as an input feature during training for random subsequences between 10 and 1000 milliseconds, and explicitly conditioning the autoregressive RNN on this ground truth F0. Explicitly conditioning the unidirectional RNN allows it to predict the correct F0 with high accuracy while, we hypothesize, the input features encourage the model to learn to generate smooth, continuous F0 surrounding the specified contour rather than suddenly jumping to the specified contour. During inference, we use the user-specified F0 values instead of the ground-truth F0. Because a recurrent model uses a combination of its input features and history and does not have a reliable history at the start of generation, discontinuities are reduced when the specified contour occurs towards the start of generation. Therefore, downstream tasks that require more editing in the second half of an utterance benefit from reversing the order of sequence generation.

Relative to DAR, C-DAR has three additional changes that do not significantly impact naturalness or controllability, but provide additional insights into F0 generation. First, a 5-layer postnet [3] follows the autoregressive RNN. We find that this postnet has the effect of reducing autoregressive sampling errors and tightening the posterior distribution around the argmax (Figure 2). Second, we use scheduled sampling [24] instead of data dropout. Scheduled sampling is known to not be a consistent estimator [25] and was shown to exhibit worse objective metrics for F0 generation (V/UV precision and pitch RMSE) [12]. Our findings indicate that neither consistency nor superior objective metrics are reliable indicators of improved subjective naturalness. Lastly, our bidirectional RNN has 16 hidden units instead of 256, indicating that prior F0 generation models may be using more model capacity than necessary.

2.3. Input features for F0 generation

For each 10 ms frame, we concatenate five input features: (1) the one-hot-encoded phoneme, (2) a BERT [26] word embed-

[1]We demonstrate this use case and provide audio examples of our experiments at `https://www.maxrmorrison.com/sites/controllable-prosody`.

ID	Gender	Book(s)	Hours
94	Male	*The Canterville Ghost*	3.00
3906	Female	*Little Fuzzy*	4.19
5717	Male	*The Cowardly Lion of Oz*	4.19
11049	Female	*The Warren Report*	7.76

Table 1: *Speakers used for evaluation and vocoder training.*

ding, (3) the V/UV label, and (4) one-hot encodings of nearby punctuation (e.g., whether the word precedes a comma or is in quotations). We use P2FA [27] for phoneme alignment. Word embeddings are computed by averaging over subword tokens extracted via the `bert-large-uncased` pretrained model from the HuggingFace Transformers package [28]. These features are jointly referred to as "Text features" in Figure 1.

2.4. F0 representation

We use a modified version of CREPE [29] to extract ground truth F0 contours. Our modification is as follows: rather than performing a localized search around the argmax of the posterior distribution over F0 bins, we directly decode the F0 contour from the posterior distribution via Viterbi decoding. Our transition matrix places maximal probability on maintaining the same F0 value and zero probability on F0 discontinuities greater than 240 cents—with linearly decreasing probability in between. We determine V/UV labels via hysteresis thresholding applied to CREPE's harmonicity confidence value. During training, we quantize our F0 representation to one of 128 values. We reserve one value for predicting unvoiced tokens and evenly divide the other 127 bins to span 4 standard deviations above and below the speaker's average F0 in base-2 log-space.

3. Pitch-shifting WaveNet vocoder

In order to synthesize speech from an arbitrary F0 contour, we propose a pitch-shifting WaveNet vocoder (PS-WaveNet) that accepts an F0 contour as conditioning. We use a publicly available implementation [30] of a single-speaker WaveNet vocoder [31] and predict 10-bit μ-law-encoded waveform samples. We enable explicit F0 control by forcing the input acoustic features through a small, jointly learned bottleneck that encourages the network to take F0 features solely from the input F0 contour.

We use 21-channel mel-cepstral coefficients (MCeps) instead of the typical 80-channel log-mel-spectrograms as MCep are less individually representative of energies at specific frequencies and therefore more easily separable from F0. Our informal experiments using 80-channel log-mel-spectrograms produced samples that were relatively inharmonic, but also captured more high-frequency detail. Our MCep bottleneck consists of three 1D convolutional layers with a filter width of 5 and ReLU activations between layers. The output channels of each convolutional layer are 20, 20, and 12, respectively. We train our PS-WaveNet vocoder on the bottlenecked MCep features concatenated with the one-hot-encoded F0 contour.

4. Experimental Design

We use the 360-hour clean training data partition from LibriTTS dataset [32] to train our F0 generation networks. We train our PS-Wavenet vocoders and evaluate all models using LJSpeech [33] as well as three single-speaker datasets similarly constructed from LibriVox recordings. The reader ID, gender, book, and amount of training data for each speaker are given in Table 1. For evaluation, we use 20 held-out utterances (10 questions and 10 statements) from each speaker with duration between 2 and 10 seconds. For all subjective listening tasks, we collect 25 responses from each of our 48 US participants on Amazon Mechanical Turk (AMT).

4.1. Model training

We train DAR and C-DAR with a batch size of 32 utterances and an ADAM optimizer [34] with a learning rate of 10^{-3}. We find that validation loss corresponds poorly with naturalness. Instead of early stopping, we train for 15 epochs (3.6k steps per epoch) and manually listen to results from the LibriTTS validation set from epochs 5-15. We find that DAR and C-DAR produce the most natural F0 contours after 9 epochs. We train one PS-WaveNet for each speaker in Table 1. We train for 475k steps with a batch size of 8 and an ADAM optimizer with a learning rate of 10^{-3} that is halved every 200k steps. We use 30 dilated convolution layers with dilation rate $2^{\ell \bmod 10}$ at layer ℓ. Noise injection with 10^{-3} Gaussian noise is used to improve the stability of synthesis [35].

4.2. Evaluating F0 contour generation

We evaluate F0 generation models using both objective and subjective metrics, but emphasize that subjective metrics align best with our goal. Our objective metrics are the pitch RMSE and the negative log-likelihood (NLL) of the model relative to ground truth pitch. We do not report V/UV metrics, as all models correctly preserve the V/UV sequence. We report two subjective metrics, including a novel subjective metric that addresses a problem with previous F0 generation evaluation methods.

Prior work on F0 generation uses pitch-shifting vocoders to generate evaluation samples, which participants listen to and provide a naturalness rating [9, 10, 11]. Here, we address the issue where artifacts induced by pitch-shifting are proportional to the size of the shift. This penalizes natural-sounding F0 contours that have high ℓ_1 or ℓ_2 distance from the original pitch, and rewards unnatural F0 contours close to the original pitch. To address this, we low-pass filter each vocoded sample at 10 Hz above the maximum F0. The resulting audio preserves the F0 contour and amplitude envelope while removing all artifacts above the cutoff frequency. During the user study, participants are told that they are listening to the intonation of speech spoken by either a real person ("real") or synthesized by a computer ("fake"), and are asked to identify each sample as real or fake.

We implement our proposed user study to evaluate the naturalness of DAR and C-DAR. We use the PSOLA vocoder, and include as baselines a monotone model as well as two random models: *replace*, which replaces the F0 contour of each word with a contour from a random word uttered by the same speaker, and *swap*, which randomly swaps F0 contours of words within the sentence. For completeness, we also conduct the more typical MOS naturalness test without low-pass filtering using our proposed vocoder (see Section 4.3).

Our second subjective study evaluates the controllability of the C-DAR model on the task of synthesizing F0 after changing a question mark to a statement, or vice versa. We call this task "repunctuation". Our weak baseline is the original audio with the original punctuation. As a strong but unnatural baseline, we replace only the last two words of a sentence with a manually-selected F0 contour that is representative of the target punctuation. For DAR and C-DAR, we change the punctuation of the text input. For C-DAR, we also provide the F0 of the last

F0 source	NLL	RMSE	% Considered Real
Original	–	0.00	0.72
Monotone	–	0.37	0.19
Random (swap)	–	0.37	0.37
Random (replace)	–	0.43	0.38
DAR	8.15	0.43	0.57
C-DAR	9.97	0.45	0.55

Table 2: *Results for objective F0 generation experiments and the subjective low-pass experiment. Lower scores are better for NLL and RMSE and higher is better for % Considered Real.*

F0 source	Vocoder	V/UV Metrics	RMSE	MOS
3-bit	None	0.99/0.64	0.09	1.48
Original	None	1.00/1.00	0.00	4.30
Original	PSOLA	0.99/0.98	0.06	4.10
Original	WORLD	0.97/0.80	0.05	3.61
Original	PS-WN	0.93/0.87	0.25	3.73
C-DAR	PSOLA	0.98/0.97	0.19	3.55
C-DAR	WORLD	0.97/0.80	0.07	3.11
C-DAR	PS-WN	0.93/0.85	0.32	3.52
DAR	PS-WN	0.94/0.84	0.28	3.41

Table 3: *Pitch-shifting vocoder experiment results. PS-WN is our proposed PS-WaveNet. V/UV metrics are formatted as* `precision/recall`.

two words as a user-specified F0 segment. Samples are vocoded using PSOLA and low-pass filtered as described above. AMT participants are given a sample and asked to select whether the sample sounds more like a statement or question.

4.3. Evaluating PS-WaveNet

We evaluate the consistency and naturalness of PS-WaveNet via two tasks. For both tasks, our baselines are PSOLA [36] and WORLD [15]—two DSP-based vocoders with frame-wise F0 control. For our first task, we measure how closely the synthesized speech follows the given F0 contour via the F0 RMSE and V/UV errors between the input and output F0. We obtain the F0 of the output using our method described in Section 2.4. For our second task, AMT participants rate the naturalness of each sample between 1 (low naturalness) and 5 (high naturalness). We evaluate all vocoders using both the original F0 contour and the F0 contour generated by C-DAR. We include the original audio and intentionally degraded audio (quantized to 3 bits) as references for high and low naturalness, respectively.

5. Results

5.1. F0 Generation

We present the F0 generation results in Table 2. We see that C-DAR achieves a comparable naturalness to DAR while enabling user control and context-awareness. This is further corroborated by the mean opinion scores (MOS) in Table 3, which show that participants considered C-DAR to be slightly more natural than DAR. The results of Table 2 also corroborate that NLL and RMSE are unsuitable metrics for F0 generation: neither correlates with subjective perceptions of naturalness. Further, we found that NLL could be trivially lowered by training C-DAR for fewer epochs, but with clearly degraded naturalness. This reinforces the need for domain-specific subjective metrics such as our proposed low-pass evaluation method.

	Original	Heuristic	Monotone	DAR
Original	-	-	-	-
Heuristic	0.55/0.54	-	-	-
Monotone	-	0.28/0.36	-	-
DAR	0.43/0.60	0.41/0.46	0.68/0.49	-
C-DAR	0.59/0.60	0.49/0.46	0.71/0.69	0.63/0.50

Table 4: *Repunctuation experiment results. A pairwise comparison of five models. All results indicate percent preference for the model specified in the same row over the model in the same column. Results are formatted as* `Q/S`*, where* `Q` *and* `S` *are the percent preferences when the target punctuations are question marks and periods, respectively.* ***Heuristic*** *is our strong baseline described in Section 4.2.*

We present our repunctuation experiment results in Table 4. Relative to DAR, using C-DAR with short, user-specified F0 contours improves the adherence of the generated F0 contour to high-level semantic concepts (e.g., questions and statements). We find this to be especially true when the target punctuation is a question mark. We believe this is because statements are heavily over-represented in the dataset, leading to class-imbalance and mode collapse. Our results indicate that simple user inputs make for an effective mode selector for prosody generation.

5.2. PS-WaveNet

In Table 3, we see that our PS-WaveNet significantly outperforms the naturalness of WORLD while achieving comparable performance to PSOLA. We find that PS-WaveNet has a higher variance of MOS across speakers, ranging from 3.04 for speaker 5717 to 3.80 for speaker 94 when using C-DAR. In comparison, PSOLA achieves 3.40 and 3.58 MOS on speakers 5717 and 94, respectively. An additional pairwise test between PS-WaveNet and PSOLA using F0 contours generated with C-DAR confirms that PSOLA is preferred only for speakers 5717 and 11049.

The objective metrics reported in Table 3 highlight additional tradeoffs when selecting a pitch-shifting method. For example, we see that WORLD achieves the best RMSE despite its low MOS, but also tends to make unvoiced regions sound voiced (i.e., low V/UV recall). This is more useful for pitch-shifting singing, for example, as high pitch accuracy is important and unvoiced regions are less common than speech. PS-WaveNet achieves higher V/UV recall than WORLD, but at a cost to V/UV precision and RMSE. We hypothesize that the increase in inharmonicity due to lower V/UV precision also induces more pitch-tracking errors, including pitch-doubling errors which produce extremely high RMSE.

6. Conclusion

In this work, we present a deep autoregressive model that supports controllable, context-aware F0 generation; a pitch-shifting neural vocoder that allows explicit F0 conditioning; and novel subjective evaluation methods for F0 generation. We show in user studies that our controllable F0 model exhibits comparable naturalness as non-controllable baselines, and that our pitch-shifting neural vocoder exhibits comparable naturalness as DSP-based vocoders. There are many directions for future work, including real-time pitch-shifting vocoding and interaction design for prosody editing.

7. References

[1] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[2] J. C. Wells, *English intonation PB and Audio CD: An introduction*. Cambridge University Press, 2006.

[3] 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.

[4] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, vol. 32, pp. 3171–3180.

[6] K. E. Dusterhoff, A. W. Black, and P. A. Taylor, "Using decision trees within the tilt intonation model to predict f0 contours." 1999.

[7] A. Raux and A. W. Black, "A unit selection approach to f0 modeling and its application to emphasis," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 700–705.

[8] K. Yu and S. Young, "Continuous f0 modeling for hmm based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2010.

[9] Z. Hodari, O. Watts, and S. King, "Using generative modelling to produce varied intonation for speech synthesis," *arXiv preprint arXiv:1906.04233*, 2019.

[10] X. Wang, S. Takaki, and J. Yamagishi, "Autoregressive neural f0 model for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1406–1419, 2018.

[11] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, "A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 157–170, 2019.

[12] X. Wang, S. Takaki, and J. Yamagishi, "An rnn-based quantized f0 model with multi-tier feedback links for text-to-speech synthesis." in *Interspeech*, 2017, pp. 1059–1063.

[13] W. Ping, K. Peng, K. Zhao, and Z. Song, "Waveflow: A compact flow-based model for raw audio," *arXiv preprint arXiv:1912.01219*, 2019.

[14] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5916–5920.

[15] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[16] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018.

[17] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[18] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. Skerry-Ryan, D. Stanton, D. Kao, and T. Bagby, "Semi-supervised generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2020.

[19] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," *arXiv preprint arXiv:2002.03785*, 2020.

[20] Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-Periodic WaveNet Vocoder: A Pitch Dependent Dilated Convolution Model for Parametric Speech Generation," in *Interspeech*, 2019, pp. 196–200.

[21] S. Im, J. Cole, and S. Baumann, "The probabilistic relationship between pitch accents and information status in public speech," in *Proceedings of Speech Prosody*, vol. 9, 2018.

[22] Y. Mass, S. Shechtman, M. Mordechay, R. Hoory, O. Shalom, G. Lev, and D. Konopnicki, "Word emphasis prediction for expressive text to speech," in *Interspeech*, 2018, pp. 2868–2872.

[23] C. Gussenhoven, "Types of focus in english," in *Topic and focus*. Springer, 2008, pp. 83–100.

[24] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[25] F. Huszár, "How (not) to train your generative model: Scheduled sampling, likelihood, adversary?" *arXiv preprint arXiv:1511.05101*, 2015.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, p. 3878, 2008.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[29] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[30] R. Yamamoto, "r9y9/wavenet_vocoder," https://github.com/r9y9/wavenet_vocoder, 2019.

[31] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *Interspeech*, 2017, pp. 1118–1122.

[32] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

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

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: a real-time speaker-dependent neural vocoder," in *The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

[36] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11. IEEE, 1986, pp. 2015–2018.