Prosodic Prominence and Boundaries in Sequence-to-Sequence Speech Synthesis

Antti Suni, Sofoklis Kakouros, Martti Vainio, Juraj Šimko

Department of Digital Humanities, University of Helsinki, Finland

firstname.secondname@helsinki.fi

Abstract

Recent advances in deep learning methods have elevated synthetic speech quality to human level, and the field is now moving towards addressing prosodic variation in synthetic speech. Despite successes in this effort, the state-of-the-art systems fall short of faithfully reproducing local prosodic events that give rise to, e.g., word-level emphasis and phrasal structure. This type of prosodic variation often reflects long-distance semantic relationships that are not accessible for end-to-end systems with a single sentence as their synthesis domain. One of the possible solutions might be conditioning the synthesized speech by explicit prosodic labels, potentially generated using longer portions of text.

In this work we evaluate whether augmenting the textual input with such prosodic labels capturing word-level prominence and phrasal boundary strength can result in more accurate realization of sentence prosody. We use an automatic wavelet-based technique to extract such labels from speech material, and use them as an input to a tacotron-like synthesis system alongside textual information.

The results of objective evaluation of synthesized speech show that using the prosodic labels significantly improves the output in terms of faithfulness of f_0 and energy contours, in comparison with state-of-the-art implementations.

Index Terms: end-to-end speech synthesis, prominence, prosodic boundaries, continuous wavelet transform

1. Introduction

In the last few years, statistical speech synthesis has undergone a major paradigm shift. Linguistic front-end (text normalization, letter-to-sound conversion, syllabification, part-of-speech tagging, phrasing, etc.) and acoustic back-end (source and filter parameters, deterministic vocoders, etc.) have been replaced by "end-to-end" systems, such as Tacotron and its derivatives [1, 2, 3]. Given a large enough corpus, the whole chain from raw text to speech (TTS) can now be jointly modelled with neural sequence-to-sequence (s2s) models, although the most successful applications in fact use elements of more traditional synthesis, e.g., separate text-normalization and text-to-phoneme conversion modules or separate (neural) vocoders. Explicit modelling of prosodic parameters like segmental durations and pitch contours has been replaced by training attention mechanism and mapping textual input to acoustic properties represented with spectrograms.

The s2s models, in particular when combined with WaveNet-style neural vocoders [4, 2], achieve quality on par with human speech, especially for isolated sentences with neutral prosody. In order to tackle more prosodically challenging tasks, neural architectures have been extended with techniques such as global style tokens [5] and vector-quantized variational autoencoders [6]. While these techniques yield impressive results in terms of modelling various global prosodic styles, they do not address prosodic variation on finer temporal scales, such as word-level emphasis and phrasal structure of the utterances. One of the reasons is the fact that this type of prosodic variation–for example emphasis associated with givenness of information–arises from long-distance dependencies in the text that falls between overall speech style and single sentence level prosody.

Another, somewhat complementary reason arises from the lack of explicit control inherent to the "black-box" machine learning architectures, such as s2s systems. On the one hand, the existing systems are not designed to capture the long-range semantic dependencies [7], on the other hand, they do not facilitate explicit control of prosody akin to older parametric synthesis approaches, where linguistic and prosodic labels were utilized and prosodic parameters were modelled separately [8].

In this paper we address the issue of explicit prosodic control by augmenting the textual material serving as an input to a s2s system with labels conveying word-level prominence and phrasal boundary strength. The labels are extracted from the training speech material using a wavelet-based technique [9]. The objective evaluation shows that this type of local prominence and phrasal boundary control significantly contributes to the faithfulness of local prosodic variation in synthesized speech.

Prosodic labelling itself has a long history in TTS. (Binary) phrase break modelling has always been a necessary component of pre-s2s synthesis systems, and word prominence augmentation has also been experimented with, most recently in [10] where acoustically labeled prominence values were applied for controlling DNN-based parametric speech synthesis. Here, the achieved prosodic control was partly negated by a decrease in perceptual quality, likely due to parametric speech representation and a small database. In the s2s paradigm, pitch accent type have been used as an additional input for Japanese synthesis [11], demonstrating improvements in both subjective and objective evaluation, but the pitch accents were annotated manually. In contrast, the present study introduces automatic prosody annotation applied to s2s synthesis using a large training corpus.

2. Methods and experiments

2.1. Prosodic Labelling

As the current synthesis models utilize tens of hours of training speech material, manual annotation of prosodic events is not a viable option. Instead, we thus use an automatic prominence annotation procedure providing word-level labels of acoustic prominence and boundary strength using a continuous wavelet transform (CWT) based method, described in full in [9].

The procedure first uses a forced-alignment of speech signal with the text. Subsequently, prosodic signals of f_0 and energy are extracted, and a word duration signal is created by plac-

ing the word duration value in a mid-point of each word and interpolating through the values. These three signals are then combined, and the combined signal (signal panel in Figure 1) is decomposed using CWT (heat map in Fig. 1).

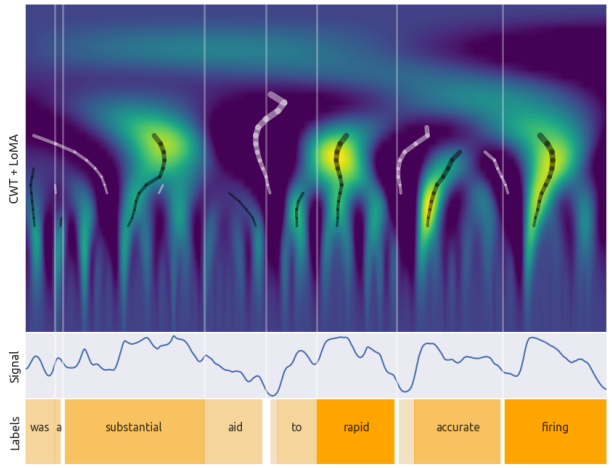

Figure 1: *CWT-based prosodic annotation method.*

CWT decomposes the signal into scales that can be associated with levels of prosodic hierarchy. To certain degree, events and movement related to, e.g., (prosodic) words and phrases can be isolated and analyzed, by following ridges or valleys across appropriate scales (black and white lines in Fig. 1). Integrating the ridge / valley lines yields continuous word prominence / boundary estimates, that can be aligned with appropriate textual units. (The continuous word level prominence and boundary strength estimates and indicated by background saturation and thickness of word boundaries, respectively, in text panel of Fig. 1).

The method is essentially unsupervised, in that no labelled data are required, yet some degree of tuning is necessary regarding (language-dependent [12]) weights of the prosodic signal types as well as discretization of resulting continuous prominence and boundary values. For the current study, the weights were tuned according to the performance on an accent and boundary detection task on Boston radio news corpus [13], for which the method achieves state-of-the-art results [9]. For word prominence estimates, the combined signal was calculated as a weighted sum of the f_0, energy and duration signals with weights 1.0, 0.5 and 1.0, respectively. For boundary strength estimates, three signals were multiplied. Both prominence and boundary values were discretized into three classes. The intervals were set manually, based on a small subset of the training utterances, such that for prominence, categories 0, 1 and 2 would roughly correspond to non-accented, accented and emphasized words. For boundaries, the phonological parallels would be no boundary, intermediate phrase boundary and intonational phrase boundary.

2.2. Sequence-to-sequence models

Sequence-to-sequence models, like Tacotron [1] are trained to generate speech spectra directly from textual input in an end-to-end fashion, with prosodic features of represented speech learned jointly alongside spectral characteristics. A front-end neural encoder encodes the textual input using a deep network combining convolutional and recurrent layers. A decoder (another stack of recurrent and convolutional layers) is trained to generate spectra in an auto-regressive, frame-by-frame manner; generation of each new frame is conditioned by a previously generated portion of the spectrogram. To provide dependency on text, the decoder is also conditioned by an output of an attention mechanism that time-aligns the output of the text encoder with the current state of the decoder network.

In the present evaluation of influence of prosodic labeling, we use a third party implementation [14] of tacotron-like architecture called Deep Convolutional TTS (DCTTS; [15]). The DCTTS model replaces the recurrent layers used in the Tacotron system with dilated convolutions and highway layers, and uses a guided attention mechanism. These design decisions alleviate the high cost of training of recurrent layers and standard attention module, and the system trains considerably faster than the original Tacotron network, without loss in output quality. See [15] for implementation and evaluation details of the system.

As in the original DCTTS implementation, the decoder module initially produces a downsampled coarse version of a MEL spectrogram that is subsequently upsampled using a Spectrogram Super-resolution Network (SSRN). We use the pre-trained SSRN from [14] for this purpose. Finally, the Griffin-Lim algorithm is used to generate an appropriate speech waveform from the full spectrogram [16].

2.3. Material

The synthesis models were trained on a large, single-speaker American English corpus LJSpeech[17], consisting of approximately 24 hours of non-fiction stories read by a professional female reader. This dataset is commonly used in deep learning speech synthesis, due to its public availability, size, and consistent, if slightly reverberant quality. Importantly for the current study, the reading style is quite expressive and the material consists of full chapters rather than isolated sentences. Informal listening reveals plenty of instances of long-range dependencies in prosody, in e.g. placing of contrastive or emphatic accents.

For prosodic labelling purposes, the dataset was aligned with Montreal forced aligner [18], using Librivox recipe. The discrete prosodic labels were then generated as described in Section 2.1, using authors' implementation of the process[1].

2.4. Implementation

To prepare the transcripts for training, the texts were phonemized (including stress marks) using CMU pronunication lexicon [19] with common punctuation (,.!?) included. Appropriate prosodic labels were simply inserted into text as additional symbols, with the word prominence labels preceding the words, and boundary labels following the words. For example, a text fragment *'I insist, that'* would (if an emphasis was detected on *insist* with a major boundary following it) be converted to:

```
<p1> ay1 <b0> <p2> ih2 n s ih1 s t , <b2> <p0> dh ae1 t
```

The dataset transcriptions augmented with prosodic labels will be available online[2].

DCTTS s2s synthesis framework [14] (with the same architecture and hyper-parameters as in the original implementation) was used to train four models differing only in the prosodic marks used to augment the input: a baseline model without prosodic labels (**dctts**), a model with both prominence and boundary labels (**P+B**) and two models with either prominence or boundary labels (**P** and **B**).

[1] https://github.com/asuni/wavelet_prosody_toolkit
[2] https://www.mv.helsinki.fi/home/asuni/sp2020/

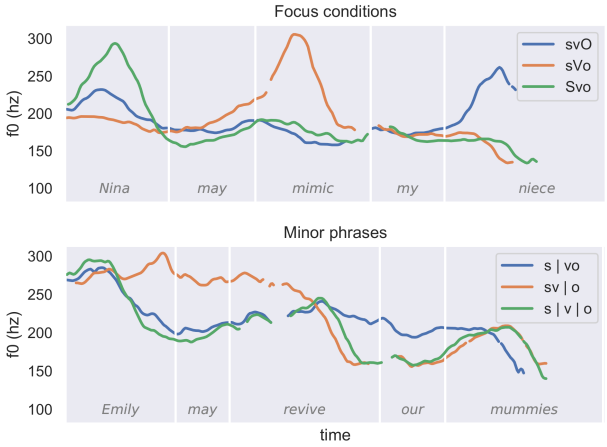

Figure 2: *Controlling focus placement and minor phrases.*

Of the 13,100 text fragments in the corpus 12,000 were used for training. Final chapter from the held-out data, 150 fragments were used for objective evaluation. The models were trained for 1000 epochs, and the test material was synthesized using oracle prosodic labels, i.e., the labels obtained for the test material in the same way as for the training set. For synthesis, the forcibly incremental attention procedure [15] was relaxed, as the default implementation appeared to generate too fast speech.

3. Evaluation

In order to evaluate the systems, we performed (1) a small qualitative assessment of prosody control, (2) a comparison between the baseline and the prosody-augmented systems, by re-labelling the prosody of synthetic test utterances and comparing the estimated labels to the reference labels, and (3) a standard objective evaluation by quantitatively comparing the prosodic signals extracted from synthetic utterances to those of the reference speech. For the objective evaluation, we included two additional TTS baselines from state-of-the-art public framework[3], a Tacotron 2 (**taco**) and Tranformer (**trans**) model, which, with more complex model structure, could hypothetically model long-range dependencies of sentence prosody better than DCTTS framework. Note, that the prosody-augmented results do not reflect realistic TTS performance, but the ideal performance of such systems; instead of predicting the labels from text, we apply the oracle labels of reference speech.

3.1. Assessment of Control

First, we informally assessed the performance of the **P+B** model and the utility of the prosodic labels for control over prosodic realization of synthetic speech.

Short subject-verb-object sentences adapted from [20] were synthesized, simulating different focus and boundary conditions by manually setting the prosodic labels in the input. The top panel in Fig. 2 shows f_0 contours of three synthesized utterances with the emphasis location (elicited by the word prominence label <p2>) on subject, verb and object, respectively. Note that the system is able to reproduce the intended foci. In the bottom panel in Fig. 2, intermediate, or minor phrases, boundaries were elicited by setting the boundary label to <b1> after subject, verb or both. Again, the three conditions are clearly differentiated, forming seemingly appropriate f_0 contours.

	dctts			**P + B**		
	prec.	rec.	F	prec.	rec.	F
prominence	acc=0.61			acc=0.81		
<p0>	0.77	0.78	0.78	0.90	0.90	0.90
<p1>	0.35	0.44	0.39	0.61	0.59	0.60
<p2>	0.57	0.42	0.48	0.79	0.80	0.80
boundary	acc=0.70			acc=0.85		
<b0>	0.78	0.82	0.80	0.88	0.92	0.90
<b1>	0.53	0.50	0.51	0.78	0.72	0.75
<b2>	0.52	0.43	0.47	0.86	0.84	0.85

Table 1: *Synthetic vs original prosodic labels*

3.2. Categorical Evaluation

Test samples were synthesized by the baseline **dctts** system and the prosodically augmented **P+B** system. For the latter, we used the oracle prosodic labels obtained from the original waveforms. The synthesised test samples were then prosodically labeled using the same procedure (Section 2.1), and the resulting labels were compared with the original ones.

Table 1 summarizes the results of the comparison for the two systems. Accuracy, precision, recall, and the F-values are higher for the prosodically augmented **P+B** system than for the baseline. The differences are greater for the emphatically accented (2) and the major boundary labels (2) than for to the unaccented and no-boundary labels (0 and 1), respectively. This indicates, that while the baseline system performed well above chance in producing distinguishable patterns of binary prominence and boundary in correct locations, it did not capture the finer distinctions of weak and strong prominence.

The **P+B** system struggled most in distinguishing the middle categories, which is expected due to somewhat arbitrary discretization of the inherently continuous prominence and boundary values.

3.3. Objective Evaluation

A set of objective measures comparing the original (reference) and the synthesized signals was used to formally evaluate the performance of the synthesis models. Although objective measures do not directly correlate with subjective measures of human perception, they provide the means to assess the overall model performance (see, e.g. [21, 22]).

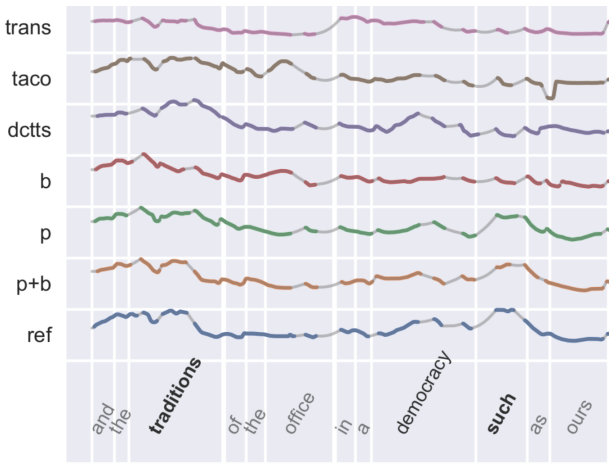

Figure 3: *Comparison of methods in f_0 generation, see text.*

The objective measures used in the current setup are the root-mean-squared error (RMSE) and Pearson correlation between the reference and the synthesized signal in terms of (i) the f_0 over the voiced intervals, (ii) the voiced energy, (iii) phone duration, and (iv) word duration.

To account for the mismatch in the time-alignment between the original and synthesized signals, the waveforms (original-synthesized) were compared using dynamic time warping (DTW) [23], and the respective features were time-aligned to match the minimum distance score across the sequences. For the computation of the word and phone metrics, segmental and word durations were extracted by forced alignment using the Montreal Forced Aligner [18]. Note that these steps (DTW, f_0 extraction and alignment) introduce some additional noise to the measurements. Fig. 3 shows examples of f_0 contours of the utterances synthesized by the evaluated s2s systems, time aligned with respect to the original reference rendition (ref); the word-shading and boundary thickness reflect oracle labels.

	P+B	P	B	dctts	taco	trans	
f_0	2.132	2.339	2.507	2.639	2.954	2.865	RMSE
energy	3.245	3.240	3.315	3.642	3.456	3.359	
ph.dur	0.025	0.025	0.026	0.027	0.027	0.029	
wd.dur	0.052	0.056	0.056	0.060	0.064	0.070	
f_0	0.655	0.595	0.519	0.471	0.460	0.457	Correl.
energy	0.677	0.661	0.653	0.605	0.620	0.627	
ph.dur	0.833	0.825	0.825	0.798	0.788	0.770	
wd.dur	0.978	0.975	0.974	0.970	0.967	0.957	

Table 2: *RMSE and correlation values for the evaluated s2s systems for f_0, energy, phone duration, and word duration. Units: semitones for f_0, spl for energy, and seconds for durations.*

Table 2 lists the mean RMSE and correlation coefficients for the values calculated for individual sentences in the test corpus, separately for comparisons between the reference and the utterances generated by different tested s2s systems. One-way anova (with a Bonferroni adjustment to compensate for multiple comparisons) was used to compare the values for different systems. The shaded cells in Table 2 indicate the attributes for which the prosody-augmented systems yielded significantly lower RMSE / higher correlation coefficient than the non-augmented system **dctts** (darker shade: $p < 0.001$, lighter shading: $p < 0.01$). As can be seen, the augmented systems performed significantly better for f_0 in terms of RMSE and correlation (except for **B** for the latter), and for energy (except RMSE for the **B** system). The prosody augmented systems also provide significantly higher correlations (but not lower RMSEs) between the aligned synthesized and reference signals in terms of phone duration. The system using both prosodic labels also reproduces the original word duration significantly better than the baseline.

For the great majority of the assessed measures, the **P+B** achieved the lowest mean RMSEs and the highest mean correlations of all evaluated systems. This performance advantage is particularly strong for f_0 measure. As shown in Fig. 4, for f_0 **P+B** in fact performs significantly better than any other of the evaluated systems. In general, both prominence **P** and boundary **B** labels improve upon baseline, and the effects are independent of each other, as combining both label types yields further improvements in most measures. Neither **taco** nor **trans** TTS model yield improvements upon **dctts** baseline, in fact their matching of the reference f_0 is quite poor, despite high perceptual quality of those systems [3].

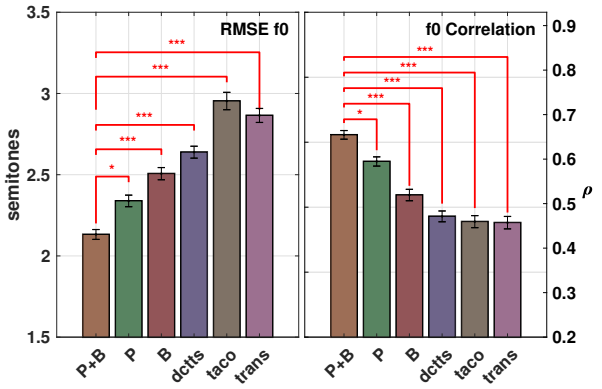

Figure 4: *Mean RMSE and correlation coefficients for voiced f_0 for all tested systems (note the truncated y-axis). Significance of the difference between the **P+B** system and all other systems is indicated in red.*

4. Discussion

The current study shows, that prosodic control and reproduction of prosodic patterns of natural speech is to a high degree achievable within a sequence-to-sequence synthesis paradigm. The next important step will be an evaluation of how do the the measurable improvements translate to perceptual quality, in particular for longer stretches of textual input.

In this work we use oracle prosodic labels extracted from the target speech material. In order to develop a fully fledged TTS system, these prosodic labels need to be predicted directly from the text. A recent contribution [24] has yielded promising results in predicting prosodic labels such as those used here from a text using contextualized word representations that can capture long-term semantic dependencies in text.

While the systems using prosodic labels reproduce signal characteristics better than the TTS baseline, the correlations with the reference speech for the prosody augmented system are still relatively low (e.g., 0.655 for f_0, etc.). It should be noted that our labeling scheme operates on slow scales of words and phrases rather than on syllables used as the fundamental unit for many prosody annotation schemes such as ToBI. While it might be unreasonable to expect that an s2s system could learn to perfectly handle local phenomena such as accent type, peak-alignment and boundary tones, the presented approach would in principle allow for a shift of the scope of prominence labels to syllable level.

Striving for such level of descriptive adequacy could, however be excessive: enforcement of too much detail could, in our opinion, hamper the ability of s2s system to generalize from the training material. Also, considerations of what can be robustly labeled from acoustics and what can be predicted from text must also be taken into account. Hence, we believe that word and phrase level labeling forms a good trade-off.

5. Acknowledgements

The work was partly funded by an Academy of Finland Research Fellowship grant (#309575) to the last author.

6. 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," *arXiv preprint arXiv:1703.10135*, 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] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," *arXiv preprint arXiv:1910.10909*, 2019.

[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," *arXiv preprint arXiv:1609.03499*, 2016.

[5] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," 2018.

[6] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[7] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," *arXiv preprint arXiv:1909.03965*, 2019.

[8] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[9] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[10] Z. Malisz, H. Berthelsen, J. Beskow, and J. Gustafson, "Promis: a statistical-parametric speech synthesis system with prominence control via a prominence network," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 257–262.

[11] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6905–6909.

[12] A. Eriksson, A. S. Suni, M. T. Vainio, J. Simko *et al.*, "The acoustic basis of lexical stress perception," in *Proceedings of the 9th International Conference on Speech Prosody 2018*. International Speech Communications Association, 2018.

[13] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[14] K. Park. (2018) A TensorFlow Implementation of DC-TTS: yet another text-to-speech model. https://github.com/Kyubyong/dc_tts. Accessed: 2019-09-29.

[15] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4784–4788.

[16] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

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

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017, pp. 498–502.

[19] R. Weide, "The carnegie mellon pronouncing dictionary [cmudict. 0.7b]," 2014.

[20] Y. Xu and C. X. Xu, "Phonetic realization of focus in english declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[21] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5140–5144.

[22] Y. Gu and Y. Kang, "Multi-task wavenet: A multi-task generative model for statistical parametric speech synthesis without fundamental frequency conditions," in *Proc. Interspeech 2018*, 2018, pp. 2007–2011. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1506

[23] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[24] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," *arXiv preprint arXiv:1908.02262*, 2019.