In Other News: A *Bi-style* Text-to-speech Model for Synthesizing Newscaster Voice with Limited Data

Nishant Prateek, Mateusz Łajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, Trevor Wood

Amazon Research Cambridge, UK

nprateek@amazon.co.uk

Abstract

Neural text-to-speech synthesis (NTTS) models have shown significant progress in generating high-quality speech, however they require a large quantity of training data. This makes creating models for multiple styles expensive and time-consuming. In this paper different styles of speech are analysed based on prosodic variations, from this a model is proposed to synthesise speech in the style of a newscaster, with just a few hours of supplementary data. We pose the problem of synthesising in a target style using limited data as that of creating a bi-style model that can synthesise both neutral-style and newscaster-style speech via a one-hot vector which factorises the two styles. We also propose conditioning the model on contextual word embeddings, and extensively evaluate it against neutral NTTS, and neutral concatenative-based synthesis. This model closes the gap in perceived style-appropriateness between natural recordings for newscaster-style of speech, and neutral speech synthesis by approximately two-thirds.

1 Introduction

Newscasters have a clearly identifiable dynamic style of speech. As more people are using virtual assistants, in their mobile devices and home appliances, for listening to daily news, synthesising newscaster-style of speech becomes commercially relevant. A newscaster-style of speech gives users a better experience when listening to news as compared to news generated in the neutral-style speech, which is typically used in text-to-speech synthesis. In addition, synthesising news using text-to-speech is more cost-effective and flexible than having to record new snippets of news with professional newscasters every time a new story breaks in.

Recent advances in neural text-to-speech (NTTS) synthesis (Van Den Oord et al., 2016; Wang et al., 2017; Shen et al., 2018; Merritt et al., 2018) have enabled researchers to generate high-quality speech with a wide range of prosodic variations. For many years, concatenative-based speech synthesis (Black and Campbell, 1995; Taylor, 2006; Qian et al., 2013; Merritt et al., 2016; Wan et al., 2017) has been the industry standard. Concatenative-based speech synthesis methods can produce high-quality speech, but are limited by the coverage of units in its database. When it comes to more expressive styles of speech, this problem is aggravated by the many hours of speech data that would be needed to cover an acceptable range of prosodic variations present in a particular style of speech. The concatenative approaches also require extensive hand-crafting of relevant low-level features, and arduous engineering efforts.

Recently proposed models based on sequence-to-sequence (seq2seq) architecture (Wang et al., 2017; Shen et al., 2018; Ping et al., 2017) attempt to alleviate some of these issues by transforming the low-level feature representation into a learning task. These models function as acoustic models which take text, in the form of characters or phonemes as input, and output low-level acoustic features that can be then converted into speech waveform using one of the several 'vocoding' techniques (Perraudin et al., 2013; Shen et al., 2018; Lorenzo-Trueba et al., 2018). Seq2seq models also allow us to condition the model on additional observed or latent attributes that help in improving the flexibility (modelling different speaker, and styles), and naturalness (Ping et al., 2017; Jia et al., 2018; Wang et al., 2018; Skerry-Ryan et al., 2018; Stanton et al., 2018). Li et. al. (2018) have explored transformer networks for context generation. This improves training efficiency while capturing long-range dependencies. Even though transformers have enabled parallel training, they still suffer from slow inference due to autoregres-

Proceedings of NAACL-HLT 2019, pages 205–213
Minneapolis, Minnesota, June 2 - June 7, 2019.

sion. LSTM-based seq2seq architectures, having lesser number of trainable parameters, allow for faster inference.

Several works have explored the "controllability" of style in synthesised speech through latent-variable modelling techniques (Akuzawa et al., 2018; Henter et al., 2018; Hsu et al., 2018). These models not only enable us to jointly model different styles, but also allow the user to control the style through modification of disentangled latent variable during the inference. Although flexible, these models usually require a large amount of data to capture the idiosyncrasies of speaking styles, and to disentangle the characteristics of speech (pitch, duration, amplitude etc.) Additionally, these models are slow to train and are potentially overly complex for modelling styles of speech that are expressive but do not display large prosodic variations. During inference, the user would need to input the latent variables to synthesise, which is not ideal for production systems.

Conventional seq2seq models for NTTS rely on a single encoder for linguistic inputs (phonemes/character embeddings). This encoder cannot be solely relied upon to capture higher-level text characteristics like syntax or semantics. The relation between syntax, semantics and prosody is complex. Many linguistic theories try to tie these phenomena but they struggle to explain some edge cases and are mutually inconsistent (Taylor, 2009) . Thus, it might be unsatisfactory to apply linguistic knowledge directly to prosody modelling by conditioning the model on manually selected features. Recent advances in representation learning for text (Peters et al., 2018; Devlin et al., 2018) have allowed us to come up with linguistic representations that not only capture the semantics of a word, but are also context-dependent as a function of the entire sentence. Contextual word embeddings (CWE) can be used to present to the model additional conditioning features that can help model the prosodic variations in each word, based on the context in which it is present.

Latorre et. al (2018) investigated the effect of data reduction on seq2seq acoustic models. They train a multispeaker model with limited data from several speakers. Chung et. al (2018) pre-train the decoder of their acoustic model on a large amount of unpaired data where the decoder learns the task of predicting the next frame. They also propose conditioning the model on traditional word-vectors like GloVe and Word2vec (Pennington et al., 2014; Mikolov et al., 2013) to provide additional linguistic information. Both these works don't look at varying prosody or speaking-style. There has been a growing interest in adaptive techniques for voice cloning (Arik et al., 2018; Chen et al., 2019), and style adaptation (Bollepalli et al., 2018) with limited data. However, these models require extensive fine-tuning. Additional investigation is needed on the performance of such adaptive models on more multi-style setting.

The contribution of this work is two-fold: (1) We propose a *'bi-style'* model that is capable of generating both a distinct newscaster style of speech, and neutral style of speech, trained only on few hours of supplementary newscaster-style data, (2) we explore the use of CWE as an additional conditioning input for prosody modelling.

2 Data Exploration

This section aims at understanding the prosodic variability in neutral-style, and newscaster-style corpora. For this purpose, we study the average variance in the natural logarithm of fundamental frequency (*lf0*) for each utterance in the two styles. The values are reported in Table 1. For contrast, we also study per-utterance *lf0* in a mixed-expressive corpus from the same speaker. We notice that among the three corpora, the neutral-style utterances have the lowest mean variance per utterance, making it more tractable and easier to model with NTTS than the other two corpora. Newscaster-style has a slightly higher mean variance given greater expressiveness, and the mixed-expressive corpus has the highest mean variance. Latent-variable models (Akuzawa et al., 2018; Hsu et al., 2018; Wang et al., 2018; Henter et al., 2018; Stanton et al., 2018) tackle the problem of modelling varied expressive corpora. As we have already discussed, these models are slow to train, and require prediction or manual injection of continuous latent variables during inference. These might not be well-suited for the task of modelling newscaster-style, which even though is expressive, has much lower mean variance per utterance than the mixed-expressive corpus.

Latorre et. al. (2018) found that a minimum of $\sim$ 15000 utterances (approximately 15 hours of data) are required to train a seq2seq acoustic model from scratch. Gathering 15 hours of data

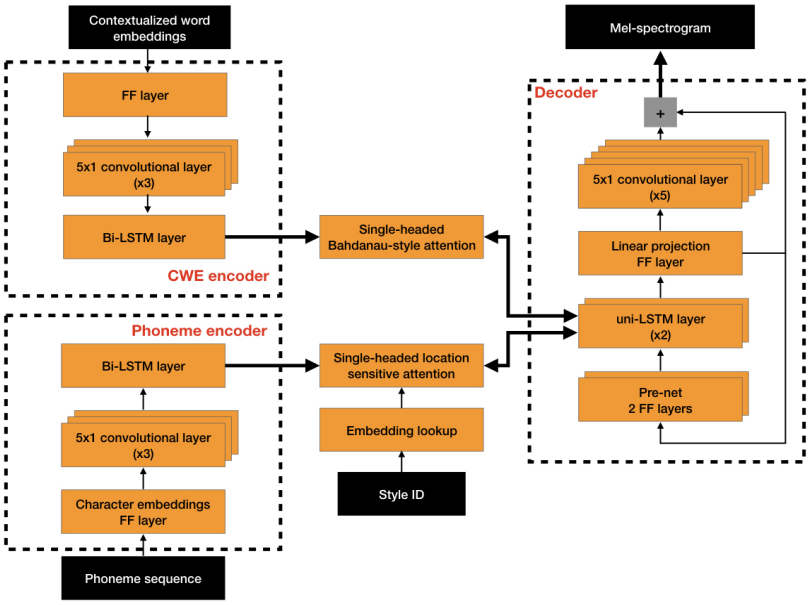

Figure 1: Context Generation Module

Corpus	Variance	Range
Neutral	6.32	5.66
Newscaster	6.33	5.68
Mixed expressive	6.79	5.71

Table 1: Analysis of mean prosodic variations based on *lf0* per utterance

for each new style is both expensive and time-consuming. Given that the mean variance for the newscaster-style utterances is marginally higher than that of neutral-style utterances, we propose jointly modelling both the neutral-style and the newscaster-style, with a one-hot *'style ID'* to differentiate between the two styles. We hypothesise that the style ID will be able to effectively factorise the neutral and newscaster styles, and generate style-appropriate samples for both. This will also alleviate the problem of prediction, and injection of continuous latent variables, that might introduce additional latency in the system. During inference, the style ID can be set by modification of simple binary flags.

From our internal corpus of female US-English voice, we use $\sim$ 20 hours of neutral-style utterances. For the newscaster-style, we use additional recordings from the same voice talent, approximating the style of American newscasters. For experiments in this paper, the amount of data used for the newscaster-style is one-fifth that of neutral-style. Using both these utterances to train a bi-style model provides us with enough overall data to train the acoustic model, and also help the model learn to factorise the two styles with the style ID input.

3 Model Description

Our proposed model is composed of two modules - Context Generation and Waveform Synthesis. The context generation module takes phonemes as inputs, and predicts temporal acoustic features, e.g. mel-spectrograms. The predicted acoustic features are then converted to time-domain audio waveforms by the Waveform Synthesis module. We provide additional inputs to the context generation module, in the form of 'style ID' and contextual word embeddings, for better prosody modelling.

3.1 Context Generation

The context generation module is an extension of the seq2seq-based acoustic model proposed by Latorre et al. (2018), and is shown in Figure 1. We propose multi-scale encoder conditioning, with the acoustic model processing phoneme-level inputs, and an additional CWE encoder that processes word-level inputs.

3.1.1 Acoustic Model

The acoustic model consists of the *phoneme encoder*, style ID input, a single-headed location-sensitive attention block, and the decoder module. The style ID is a two-dimensional one-hot vec-

tor (representing whether the input utterance belongs is in the neutral-style or newscaster-style), which is projected into continuous space by an embedding lookup layer to produce a *style embedding*. The style embedding is concatenated at each step of the output of the phoneme encoder. Single-headed location-sensitive attention (Chorowski et al., 2015) is applied to the concatenated outputs. A unidirectional LSTM-layer takes the concatenated vector of the output vector of the attention block and the pre-net layer as an input. The decoder, in each step, predicts blocks of 5 frames of 80-dimensional mel-spectrograms. We define a frame as a $50ms$ sequence, with an overlap of $12.5ms$. The last frame of the previous outputs is passed to the pre-net layer as input for generating the next set of frames.

3.1.2 CWE Encoder

We use *Embeddings from Language Models* (ELMo), introduced by Peters et al. (2018) for obtaining the contextual word embeddings for the input utterance. ELMo takes advantage of unsupervised language modelling task to learn rich text representations on a large text corpus. These representations can then be transferred to downstream tasks that often require explicit labels. ELMo embeddings bring a significant improvement for a variety of Natural Language Processing (NLP) tasks. They are able to capture both semantic and syntactic relations between words (Perone et al., 2018). As such, they seem to be a good fit for modelling prosody.

For each sentence in the training set we extract ELMo features using publicly available CLI tool (Gardner et al., 2018). This model is pre-trained on the 1 Billion Word Benchmark dataset (Chelba et al., 2014). We only use hidden states from the top layer of bi-directional Language Model (biLM). This produces a sequence of 1024-dimensional vectors, one for each word in a sentence. During training these vectors are fed to *CWE encoder*. CWE encoder has a similar topology to the phoneme encoder.

Encoded ELMo embeddings are passed to the decoder through Bahdanau-style attention (Bahdanau et al., 2015). It operates independently of location sensitive attention for phoneme encodings. It can attend to encodings of words that are not focused by location sensitive attention. We hypothesise that this can help the decoder to consider a broader context.

3.2 Waveform Synthesis

We use the pre-trained speaker-independent RNN-based "neural vocoder" proposed by Lorenzo-Trueba et. al. (2018) to convert the mel-spectrograms predicted by our context generation module into high-fidelity audio waveforms.

4 Experimental Protocol

4.1 Training

The news stories are on an average longer than neutral-style utterances, and consist of multiple sentences. Seq2seq models have a tendency to lose attention and have misalignment in longer input sequences during inference. To alleviate this, we split the news stories into individual sentences in both the training and the test sets. Splitting into individual sentences also enables us to train the model on larger batch size, helping the model to converge faster and with lesser perturbation of the training loss. To convert the utterances into phoneme sequences, we use our internal grapheme-to-phone mapping tool, which encodes the phonemes, stress marks, and punctuations as one-hot vectors.

We train the model using an L1 loss in the decoder output for mel-spectrogram prediction. To indicate when to stop predicting the decoder outputs, we have a linear stop token generator at the decoder outputs, trained jointly with the context generation module. The stop token generator is trained with an L2 loss. During training, the stop token is linearly increased from 0 at the beginning of the sentence to 1 at the end.

ADAM optimizer (Kingma and Ba, 2014) is used to minimise the training loss, with learning rate decay. The model is trained with teacher-forcing on the decoder outputs. The attention weights are normalised to add up to 1 using a softmax layer.

We use mel-spectrogram distortion (Kubichek, 1993) to monitor the input-output alignment, and the training loss to get a rough estimation on the convergence of our model. We also synthesise some held-out sentences to monitor the segmental quality and the prosody of our system, as the perceptual quality of the generated samples does not always align with the lower training and validation losses, and spectrogram distortion metrics.

System	Description
Concatenative	Concatenative-based unit selection system driven by state-level statistical parametric predictions
Neutral	Neutral-style NTTS speech
News w/o CWE	Newscaster-style NTTS speech without CWE conditioning
News with CWE	Newscaster-style NTTS speech with CWE conditioning
Recordings	Natural speech waveforms

Table 2: Systems present in the MUSHRA evaluation

4.2 Setup for Evaluation

4.2.1 Objective Metrics

We compare acoustic parameters extracted from the synthesised sentences, and the natural recordings for the analysis of prosody and segmental quality. To match the predicted sequence length to the reference sequence length for all comparisons, we use the dynamic time warping (DTW) algorithm (Bellman and Kalaba, 1959).

We use Mel-spectrogram Distortion to assess the segmental quality of the synthesised sentences.

Mel-spectrogram distortion (MSD) (Kubichek, 1993) measures the distortion between predicted and extracted (from natural speech) mel-spectrogram coefficients and is defined as:

$$MSD = \frac{\alpha}{T}\sum_{t=1}^{T}\sqrt{\sum_{d=1}^{D-1}(c_d(t)-\hat{c}_d(t))^2} \quad (1)$$

$$\alpha = \frac{10\sqrt{2}}{ln10} \quad (2)$$

where $c_d(t)$, $\hat{c}_d(t)$ are the d-th mel-spectrogram coefficient of the t-th frame from reference and predicted. T denotes the total number of frames in each utterance and D is the dimensionality of the mel-spectrogram coefficients. For our experiments, we use 80 coefficients per speech frame. The zeroth coefficient (overall energy) is excluded from MSD computation, as shown in equation 1.

For evaluating prosody, we use the following metrics calculated on *lf0*:

F0 Root Mean Square Error (FRMSE) is defined as:

$$FRMSE = \sqrt{\frac{\sum_{t=1}^{T}(x_t-\hat{x}_t)^2}{T}} \quad (3)$$

where x_t and $\hat{x}_t$ in our work denote *lf0* extracted from reference and predicted audio respectively.

F0 Linear Correlation Coefficient (FCORR) is the measure of the direct linear relationship between the predicted *lf0* and the reference *lf0*. It is expressed as:

$$\frac{T\sum(x_t\hat{x}_t)-(\sum x_t)(\sum \hat{x}_t)}{\sqrt{T(\sum x_t^2)-(\sum x_t)^2}\sqrt{T(\sum \hat{x}_t{}^2)-(\sum \hat{x}_t)^2}} \quad (4)$$

If x_t and $\hat{x}_t$ have a strong positive linear correlation, FCORR is close to +1.

Gross pitch error (GPE) (Nakatani et al., 2008) is measured as percentage of voiced frames whose relative *lf0* error is more than 20%. Relative *lf0* error is defined as:

$$\frac{|x_t-\hat{x}_t|}{x_t}\times 100 \quad (5)$$

Fine pitch error (FPE) (Krubsack and Niederjohn, 1991) is measured as standard deviation of the distribution of relative *lf0* errors, for which relative *lf0* error is less than 20%.

Since we don't explicitly predict *lf0*, we use *lf0* extracted from natural recordings, and synthesised sentences for computation of the objective metrics described above.

4.2.2 Subjective Evaluations

Even though the objective metrics give us a general indication on the prosody and segmental quality of synthesised speech, the metrics may not directly correlate to the perceptual quality. We conduct additional subjective evaluations with human listeners and consider these as the final outcome of our experiments.

For subjective evaluations, we concatenate the synthesised news-style sentences into full news stories, to capture the overall experience of our intended use-case. Each utterance is 3-5 sentences long, and the average duration is $33.47 seconds$. We test our system with 10 expert listeners with native linguistic proficiency in English, using the

MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) methodology (ITUR Recommendation, 2001). The systems used in this evaluation are described in Table 2. The listeners are asked to rate the appropriateness of each system as a newscaster voice on a scale of 0 to 100. For each utterance, 5 stimuli are presented to the listeners side-by-side on the same screen, representing the 5 test systems in a random order. Each listener rates 51 screens.

5 Results

5.1 Analysis of Objective Metrics

The scores for the objective metrics are shown in Table 3. We observe that both of our newscaster-style models obtain consistently better scores on all metrics, than neutral NTTS and concatenative-based system. Furthermore, we also observe that conditioning the newscaster-style model with CWE helps improve the prosody of the synthesised utterances.
There's a slight loss in segmental quality when conditioning the model with CWE, but it appears to be imperceptible to human listeners.

5.2 Analysis of MUSHRA Scores

The listener responses from the subjective evaluation are shown in Figure 2. In Table 4 the descriptive statistics for the MUSHRA evaluation are reported. The proposed model closes the gap between concatenative-based synthesis for newsreading, which is still largely the industry standard, and the natural recordings by 69.7%. The gap compared with the neutral NTTS voice is also closed by 60.9%. All of the systems present in the MUSHRA test are statistically significant from each other at a p-value of 0.01. This significance is observed across the listener responses using a t-test. Holm-Bonferroni correction was applied due to the number of condition pairs to compare. This significance is also observed over the MUSHRA responses in terms of the rank order awarded by listeners. For this a Wilcoxon signed-rank test applying Holm-Bonferroni correction was used.

The concatenative-based system is prone to audible artefacts at the concatenation-points, primarily due to abrupt changes in fundamental frequency in voiced phonemes. This reduces the perceived naturalness of synthesised speech. The neutral-style system is unable to model the prosody that is distinct to the newscaster-style of speech. A higher score for the newscaster-style model with CWE conditioning with respect to the model without, provides evidence supporting the hypothesis that we made in Section 1 that CWE features help model the prosodic variation better given the additional information on the syntactic context of words in the sentence.

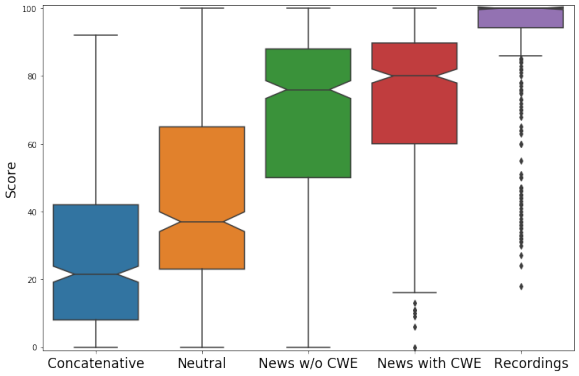

Figure 2: Boxplot of the listener responses in the MUSHRA evaluation

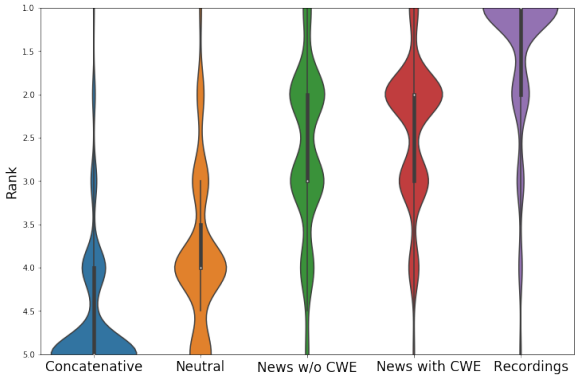

Figure 3: Violin plot of the rank-order awarded by listeners

We also generated a violin plot (Figure 3) depicting the distribution of the rank-order awarded to the systems in the test. We notice that for some of the utterances, the listeners have ranked our newsreader voice (both with and without CWE) higher than the natural recordings, showing that our context generation module is able to closely mimic the recordings in terms of prosody and naturalness.

5.3 Effect of Contextual Word Embeddings on Prosody Modelling

To further reinforce the effect of CWE on prosody modelling for newscaster-style, a preference test was conducted comparing newscaster-style with and without CWE conditioning, using 10 expert

	Segmental Quality	Prosody			
System	**MSD (dB)**	**FRMSE (Hz)**	**FCORR**	**GPE (%)**	**FPE (cents)**
Concatenative	6.07	44.85	0.28	33.58	5.68
Neutral	5.27	44.81	0.30	32.02	5.63
News w/o CWE	**4.52**	42.90	0.35	28.89	5.57
News with CWE	4.54	**42.14**	**0.36**	**27.59**	**5.55**

Table 3: Objective metrics for analysis of prosody and segmental quality. High FCORR indicates better prosody. For all other metrics, lower value indicates better performance.

System	Mean score	Median score	Mean Rank	Median Rank
Concatenative	28.31	21.5	4.60	5
Neutral	42.44	37.0	3.86	4
News w/o CWE	68.15	76.0	2.67	3
News with CWE	72.4	80.0	2.41	2
Recordings	91.61	100.0	1.45	1

Table 4: Listener ratings from the MUSHRA evaluation

listeners. Listeners were informed to rate the systems in terms of their naturalness, and were asked to choose between News with CWE, News w/o CWE, or indicate *No Preference*(NP).

Preference	Votes
News with CWE	43.2%
News w/o CWE	31%
No Preference	25.8%

Table 5: Preference test between systems with and without CWE conditioning

The listener responses are shown in Table 5. The samples conditioned on contextual word embeddings are shown to be significantly preferred (43.2%) over the samples generated without (31%), with $p < 0.01$. A binomial test was used to detect statistical significance.

5.4 Analysis of Speech Tempo

We define speech tempo of a corpus as the average number of phonemes present per second. Speech tempo is a crucial aspect in differentiating between the neutral and the newscaster styles. The newscaster-style is more dynamic than the neutral-style utterances, with higher speech tempo. In Table 6 we report the speech tempo in the neutral-style, and the newscaster-style for natural recordings, and compare those with our models with and without CWE. We observe that the model conditioned on CWE can better model the speech tempo in both styles. This gives us additional evidence that conditioning the model on CWE helps us synthesise samples that are not only more style-appropriate, but are also better in naturalness with respect to natural recordings. Analysis of speech tempo also shows us that the model is able to factorise, and replicate during inference, both styles using just a one-hot style ID.

System	Neutral	Newscaster
Recordings	11.63	14.02
with CWE	10.12	13.88
w/o CWE	10.11	13.65

Table 6: Speech tempo: recordings vs test systems

6 Conclusions

We proposed a bi-style model for generating neutral and newscaster styles of speech. We also proposed multi-scale encoder conditioning, focusing on phoneme-level and word-level inputs. Our proposed model is shown to be able to generate high-quality newsreader voice, which is significantly preferred over the neutral-style voice. We showed that the two styles can be factorised using a one-hot style ID. We also showed that the introduction of CWE conditioning significantly improves the prosody modelling ability of our context generation module, and hope that this result inspires more research into the use of NLP features in NTTS.

References

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. In *Interspeech*, pages 3067–3071.

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, pages 10040–10050.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Richard Bellman and Robert Kalaba. 1959. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9.

Alan W Black and Nick Campbell. 1995. Optimising selection of units from speech databases for concatenative synthesis.

Bajibabu Bollepalli, Lauri Juvela, and Paavo Alku. 2018. Speaking style adaptation in text-to-speech synthesis using sequence-to-sequence models with attention. *CoRR*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, and Qi Ge. 2014. orsten brants, phillipp koehn, and tony robinson. 2014. one billion word benchmark for measuring progress in statistical language modeling. INTERSPEECH.

Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Caglar Gulcehre, Aäron van den Oord, Oriol Vinyals, and Nando de Freitas. 2019. Sample efficient adaptive text-to-speech. In *International Conference on Learning Representations*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan. 2018. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. *CoRR*, abs/1808.10128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *ACL workshop for NLP Open Source Software*.

Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. 2018. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. *CoRR*, abs/1807.11470.

Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. 2018. Hierarchical generative modeling for controllable speech synthesis. *CoRR*, abs/1810.07217.

ITUR Recommendation. 2001. Method for the subjective assessment of intermediate sound quality (mushra). *ITU, BS*, pages 1543–1.

Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*.

Diederik P Kingma and Jimmy Lei Ba. 2014. Adam: Amethod for stochastic optimization. In *Proc. 3rd Int. Conf. Learn. Representations*.

David A Krubsack and Russell J Niederjohn. 1991. An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech. *IEEE Transactions on signal processing*, 39(2):319–329.

R Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128. IEEE.

Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman, Srikanth Ronanki, and Viacheslav Klimkov. 2018. Effect of data reduction on sequence-to-sequence neural tts. *CoRR*, abs/1811.06315.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Xiukun Lin, and Ming Zhou. 2018. Close to human quality tts with transformer. *CoRR*, abs/1809.08895.

Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, and Roberto Barra-Chicote. 2018. Robust universal neural vocoding. *CoRR*, abs/1811.06292.

Thomas Merritt, Robert AJ Clark, Zhizheng Wu, Junichi Yamagishi, and Simon King. 2016. Deep neural network-guided unit selection synthesis. In *ICASSP)*, pages 5145–5149. IEEE.

Thomas Merritt, Bartosz Putrycz, Adam Nadolski, Tianjun Ye, Daniel Korzekwa, Wiktor Dolecki, Thomas Drugman, Viacheslav Klimkov, Alexis Moinet, Andrew Breen, et al. 2018. Comprehensive evaluation of statistical speech waveform synthesis. In *SLT*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomohiro Nakatani, Shigeaki Amano, Toshio Irino, Kentaro Ishizuka, and Tadahisa Kondo. 2008. A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments. *Speech Communication*, 50(3):203–214.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *CoRR*, abs/1806.06259.

Nathanaël Perraudin, Peter Balazs, and Peter L Søndergaard. 2013. A fast griffin-lim algorithm. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 1–4. IEEE.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning.

Yao Qian, Frank K Soong, and Zhi-Jie Yan. 2013. A unified trajectory tiling approach to high quality speech rendering. *IEEE transactions on audio, speech, and language processing*, 21(2):280–290.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *ICML*.

Daisy Stanton, Yuxuan Wang, and R. J. Skerry-Ryan. 2018. Predicting expressive speaking style from text in end-to-end speech synthesis. *CoRR*, abs/1808.01410.

Paul Taylor. 2006. The target cost formulation in unit selection speech synthesis. In *Ninth International Conference on Spoken Language Processing*.

Paul Taylor. 2009. *Text-to-speech synthesis*, pages 111–112. Cambridge university press.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *SSW*, page 125.

Vincent Wan, Yannis Agiomyrgiannakis, Hanna Silen, and Jakub Vit. 2017. Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders. In *Proc. Interspeech*, pages 1143–1147.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: A fully end-to-end text-to-speech synthesis model. *CoRR*, abs/1703.10135.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *ICML*.