CHAR2WAV: END-TO-END SPEECH SYNTHESIS

Jose Sotelo, Soroush Mehri, Kundan Kumar[*], **João Felipe Santos**[†], **Kyle Kastner, Aaron Courville**[‡], **Yoshua Bengio**[§]
Université de Montréal
[*] IIT Kanpur
[†] INRS-EMT
[‡] CIFAR Fellow
[§] Senior CIFAR Fellow

ABSTRACT

We present Char2Wav, an end-to-end model for speech synthesis. Char2Wav has two components: a **reader** and a **neural vocoder**. The *reader* is an encoder-decoder model with attention. The encoder is a bidirectional recurrent neural network that accepts text or phonemes as inputs, while the decoder is a recurrent neural network (RNN) with attention that produces vocoder acoustic features. *Neural vocoder* refers to a conditional extension of SampleRNN which generates raw waveform samples from intermediate representations. Unlike traditional models for speech synthesis, Char2Wav learns to produce audio directly from text.

1 INTRODUCTION

The main task in speech synthesis consists of mapping text to audio signal. There are two primary goals in speech synthesis: **intelligibility** and **naturalness**. *Intelligibility* describes the clarity of the synthesized audio, specifically how well a listener is able to extract the original message. *Naturalness* describes information not directly captured by intelligibility, such as overall ease of listening, global stylistic consistency, regional or language level nuances, among others.

With traditional speech synthesis approaches, this task has been accomplished by dividing the problem into two stages. The first stage, known as the **frontend**, transforms the text into linguistic features. These linguistic features usually include phone, syllable, word, phrase and utterance-level features (Zen, 2006; Zen et al., 2013; van den Oord et al., 2016). The second stage, known as the **backend**, takes as input the linguistic features generated by the frontend and produces the corresponding sound. WaveNets (van den Oord et al., 2016) are a high quality approach to a "neural backend". For a more detailed review of traditional models for speech synthesis, we recommend consulting Taylor (2009).

Defining good linguistic features is often time-consuming and language specific. In this paper, we integrate the frontend and the backend and learn the whole process end-to-end. This procedure eliminates the need for expert linguistic knowledge, which removes a major bottleneck in creating synthesizers for new languages. We use a powerful model to learn this information from the data.

2 RELATED WORK

Attention based models have been previously used in machine translation (Cho et al., 2014; Bahdanau et al., 2015), speech recognition (Chorowski et al., 2015; Chan et al., 2016), and computer vision Xu et al. (2015) among other applications. Our work has been heavily influenced by the work of Alex Graves (Graves, 2013; 2015). In a guest lecture Graves demonstrated a speech synthesis model using an attention mechanism, an extension of his previous work on handwriting generation. Unfortunately, the speech extension was never published, so we cannot directly compare our approach to his work. However, his results were a key inspiration to us, and we hope that this work can be useful as a starting point for further developments in end-to-end speech synthesis.

3 MODEL DESCRIPTION

3.1 READER

We adopt the notation of Chorowski et al. (2015). An attention-based recurrent sequence generator (ARSG) is a recurrent neural network that generates a sequence $Y = (y_1, \ldots, y_T)$ conditioned on an input sequence X. X is preprocessed by an encoder that outputs a sequence $h = (h_1, \ldots, h_L)$. In this work, the output Y is a sequence of acoustic features and X is the text or the phoneme sequence to be generated. Furthermore, the encoder is a bidirectional recurrent network.

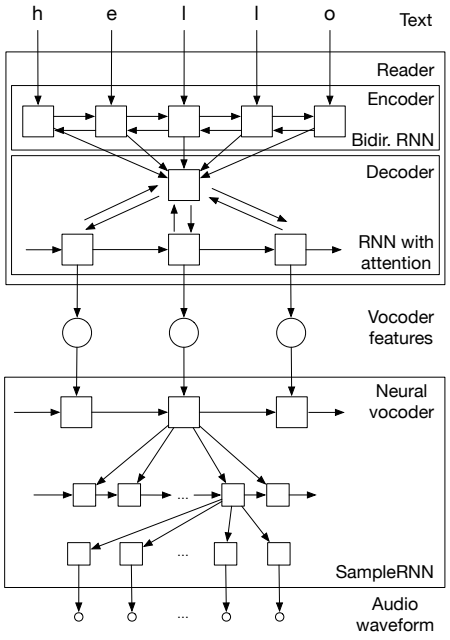

Figure 1: Char2Wav: An end-to-end speech synthesis model.

At the i-th step the ARSG focuses on h and generates y_i:

$$\alpha_i = Attend(s_{i-1}, \alpha_{i-1}, h) \tag{1}$$

$$g_i = \sum_{j=1}^{L} \alpha_{i,j} h_j \tag{2}$$

$$y_i \sim Generate(s_{i-1}, g_i) \tag{3}$$

$$s_i = RNN(s_{i-1}, g_i, y_i) \tag{4}$$

where s_{i-1} is the $(i-1)$-th state of the **generator** recurrent neural network and $\alpha_i \in \mathcal{R}^L$ are the attention weights or alignment.

In this work, we use the location-based attention mechanism developed by Graves (2013). We have $\alpha_i = Attend(s_{i-1}, \alpha_{i-1})$ and given a length L conditioning sequence h, we have:

$$\phi(i, l) = \sum_{k=1}^{K} \rho_i^k \exp(-\beta_i^k (\kappa_i^k - l)^2) \tag{5}$$

$$\alpha_i = \sum_{l=1}^{L} \phi(i, l) \tag{6}$$

where κ_i, β_i, and ρ_i represent the location, width and importance of the window respectively.

3.2 NEURAL VOCODER

Speech generation using a vocoder is limited by the reconstruction quality of that specific vocoder. To enable high quality output, we replace the vocoder with a learned parametric neural module. We use SampleRNN (Mehri et al., 2016) as an enhanced function approximator for this purpose.

SampleRNN has recently been proposed to model extremely long-term dependencies in sequential data such as audio signals. The hierarchical structure in SampleRNN is designed to capture dynamics of a sequence at different time scales. This is necessary to capture long range correlations between distant audio timesteps (e.g. word-level correlations in speech signals) as well as nearby audio timesteps dynamics.

We use a conditional version of the same model to learn the mapping from a sequence of vocoder features to corresponding audio samples. Each vocoder feature frame is added as an extra input to the corresponding state in the top tier. This allows the module to use the past audio samples and vocoder feature frames to generate the current audio samples.

4 TRAINING DETAILS

First, we pretrained the *reader* and the *neural vocoder* separately. We used normalized WORLD vocoder features (Morise et al., 2016; Wu et al., 2016) as targets for the *reader* and as inputs for the *neural vocoder*. Finally, we fine-tuned the whole model end-to-end. Our code is available online.[1]

5 RESULTS

We do not provide a comprehensive quantitative analysis of results at this time. Instead, we provide samples from our model.[2]. In Figure 2, we show samples generated by our model and their corresponding alignments to the text.

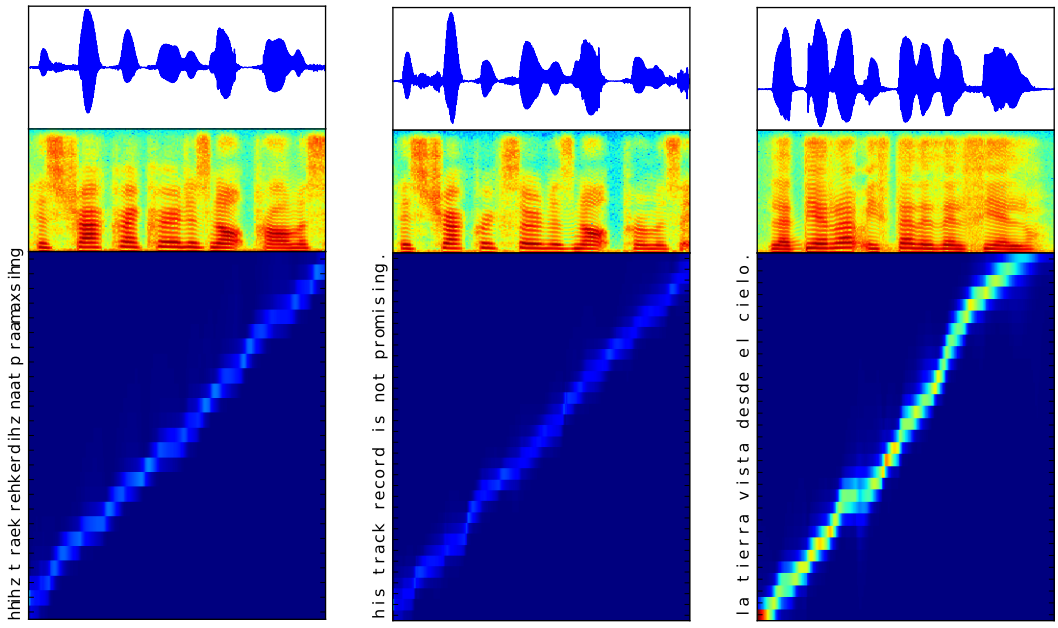

Figure 2: Samples from the models conditioned on a) English phonemes, b) English text and c) Spanish text. The models for a) and b) were trained on the VCTK dataset (Yamagishi, 2012) whereas the model for c) was trained on the DIMEX-100 dataset (Pineda et al., 2010).

ACKNOWLEDGMENTS

We thank the developers of Merlin (Wu et al., 2016), Theano (Theano Development Team, 2016), and Blocks (van Merriënboer et al., 2015). This work used the computational resources provided by Compute Canada and Calcul Québec. João Felipe Santos and Jose Sotelo thank the Fonds de Recherche du Québec - Nature et Technologies (FQRNT) for their support. Jose Sotelo also thanks the Consejo Nacional de Ciencia y Tecnología (CONACyT) as well as the Secretaría de Educación Pública (SEP) for their support. We thank the patient explanations of Dzmitry Bahdanau. His comments greatly improved this work. We also thank Alexandre de Brébisson, Jonathan Lucuix-André, Laurent Dinh, Ishaan Gulrajani, Ritesh Kumar, David Warde-Farley, Eugene Belilovsky, Natasha Jacques, Tim Cooijmans, Anna Huang, and Junyoung Chung for their comments and suggestions.

[1] http://github.com/sotelo/parrot
[2] http://josesotelo.com/speechsynthesis

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Jacob Benesty, M Mohan Sondhi, and Yiteng Huang. *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1171–1179. Curran Associates, Inc., 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003. ISSN 1532-4435.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, March 2016.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 577–585. Curran Associates, Inc., 2015.

Yuchen Fan, Yao Qian, Frank K. Soong, and Lei He. Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4475–4479, April 2015.

Alex Graves. Practical variational inference for neural networks. In J. Shawe-taylor, R.s. Zemel, P. Bartlett, F.c.n. Pereira, and K.q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2348–2356. 2011.

Alex Graves. Generating sequences with recurrent neural networks. 08 2013. URL `https://arxiv.org/abs/1308.0850`.

Alex Graves. Hallucination with recurrent neural networks, 2015. URL `https://www.youtube.com/watch?v=-yX1SYeDHbg`.

Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In Tony Jebara and Eric P. Xing (eds.), *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1764–1772. JMLR Workshop and Conference Proceedings, 2014.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. 10 2014. URL `https://arxiv.org/abs/1410.5401`.

Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pp. 373–376. IEEE, 1996.

Zeyu Jin, Adam Finkelstein, Stephen DiVerdi, Jingwan Lu, and Gautham J Mysore. Cute: A concatenative method for voice conversion using exemplar-based unit selection. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 5660–5664. IEEE, 2016.

Alexander Rosenberg Johansen, Jonas Meinertz Hansen, Elias Khazen Obeid, Casper Kaae Sønderby, and Ole Winther. Neural machine translation with characters and hierarchical encoding. 10 2016. URL https://arxiv.org/abs/1610.06550.

Geoffrey Zweig Kaisheng Yao. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. ISCA - International Speech Communication Association, May 2015.

Simon King. Measuring a decade of progress in text-to-speech. *Loquens*, 1(1), 1 2014. ISSN 2386-2637.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Bo Li and Heiga Zen. Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis. 2016.

Zhen-Hua Ling, Shiyin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen Meng, and Li Deng. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, 32:35–52, 2015.

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. 12 2016. URL https://arxiv.org/abs/1612.07837.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2204–2212. Curran Associates, Inc., 2014.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99.D(7):1877–1884, 2016.

Luis A. Pineda, Hayde Castellanos, Javier Cuétara, Lucian Galescu, Janet Juárez, Joaquim Llisterri, Patricia Pérez, and Luis Villaseñor. The corpus dimex100: Transcription and evaluation. *Lang. Resour. Eval.*, 44(4):347–370, December 2010.

Kanishka Rao, Fuchun Peng, Hasim Sak, and Francoise Beaufays. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4225–4229, April 2015. doi: 10.1109/ICASSP.2015.7178767.

Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014.

Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge, 2009.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688.

Keiichi Tokuda and Heiga Zen. Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4215–4219, 2015.

Keiichi Tokuda and Heiga Zen. Directly modeling voiced and unvoiced components in speech waveforms by neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5640–5644, 2016.

Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5): 1234–1252, May 2013. ISSN 0018-9219.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. 09 2016. URL https://arxiv.org/abs/1609.03499.

Bart van Merriënboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. Blocks and fuel: Frameworks for deep learning. *CoRR*, abs/1506.00619, 2015. URL http://arxiv.org/abs/1506.00619.

Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, and Simon King. A study of speaker adaptation for dnn-based speech synthesis. In *INTERSPEECH*, pp. 879–883. ISCA, 2015.

Zhizheng Wu, Oliver Watts, and Simon King. *Merlin: An Open Source Neural Network Speech Synthesis System*. 7 2016.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In David Blei and Francis Bach (eds.), *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 2048–2057. JMLR Workshop and Conference Proceedings, 2015.

Junichi Yamagishi. English multi-speaker corpus for cstr voice cloning toolkit, 2012. URL http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.

Heiga Zen. An example of context-dependent label format for hmm-based speech synthesis in english, 2006. URL http://hts.sp.nitech.ac.jp/?Download.

Heiga Zen. Acoustic modeling in statistical parametric speech synthesis - from hmm to lstm-rnn. In *Proc. MLSLP*, 2015. Invited paper.

Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7962–7966, 2013.

Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, and Przemysław Szczepaniak. Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices. In *Proc. Interspeech*, San Francisco, CA, USA, 2016.