Wavebender GAN: Controllable speech synthesis for speech sciences

Gustavo Teodoro Döhler Beck[1], Ulme Wennberg[1],
Gustav Eje Henter[1], Zofia Malisz[1]
[1]*Speech, Music, and Hearing, KTH Royal Institute of Technology, Sweden*

Artificial modeling of human speech has depended on an ongoing dialogue between speech scientists and engineers: speech science helped synthesis get started [1]. Reciprocally, insights into speech sciences, such as evidence for categorical speech perception and speech perception theory were reached with the use of synthetic stimuli [2].

Unfortunately, the fields have grown apart in pursuit of different goals. Speech technology has strived for ever more realistic-sounding synthesis, recently culminating in neural vocoders [3] and sequence-to-sequence systems [4] based on deep learning. The result being that technology is now capable of imitating human speech remarkably well [5], but with little or no explicit control over the output. Synthesis controlability, i.e., the ability to create and manipulate stimuli with precise control over acoustic cues such as pitch, duration, etc., is central to speech-research goals. Particularly those involving the disentanglement of different types of information in speech and their perceptual and neurophysiological correlates.

As speech technology has as of yet been unable to offer adequate control functionality, speech science remains reliant on outdated synthesis methods such as formant-based speech generation [6] (1950s) or acoustic feature editing (e.g., PSOLA [7], 1990s) that do offer control functionality. However, as these methods generally have low perceptual similarity to natural speech, the field runs the risk of insufficient universality and robustness of the associated research findings [8].

This work aims to push for progress in both speech science and technology by combining synthesis realism and control. We propose Wavebender GAN, a speech-synthesis system capable of bridging this gap. The idea of Wavebender GAN is to use deep learning to predict mel-spectrograms from low-level signal properties alone (e.g., formants, spectral slope, and the f0 contour). A high-quality speech waveform is then synthesized using state-of-the-art neural vocoders such as WaveGlow [9] and HiFi-GAN [10]. This resembles some modern text-to-speech systems with f0 control [11, 12], but we use low-level signal properties as the only inputs, and no text.

At a glance (see Figure 1), the process of creating our Wavebender GAN system can be split into four key stages: (1) select uncorrelated low-level signal properties as system inputs, that contain sufficient information to predict natural-sounding speech mel-spectrograms and are also of interest to manipulate independently; (2) perform data augmentation (e.g., pitch and gain manipulation) to allow the model to explore the domain of relevant input and output features more effectively; (3) use the augmented data to train *Wavebender Net*, a version of the ResNet architecture [13] adapted to predict mel-spectrograms from the selected low-level signal properties; and (4) improve the realism of the predicted mel-spectrograms by enhancing them using a conditional GAN (cGAN) [14]. Currently, our Wavebender GAN has been trained on the publicly available, single-speaker LJ Speech dataset [15]. Initial subjective evaluations of output control and quality suggest very good results on both measures.

Taken together, Wavebender GAN enables speech scientists to construct end-to-end pipelines for stimulus creation and testing of phonological models. The system provides a technological update to these pipelines in that it generates synthetic speech signals that are controllable as well as correlated with a larger share of natural speech cues.

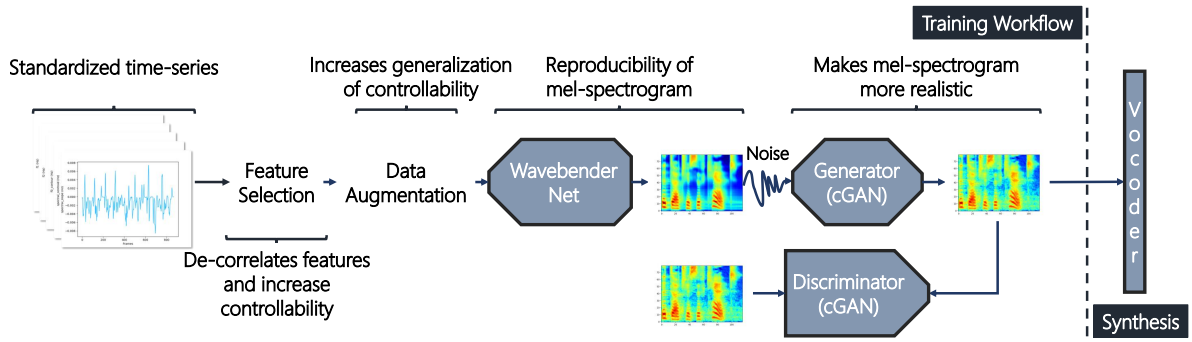

Figure 1: Wavebender GAN implementation workflow.

References

[1] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, p. e006, 2014.

[2] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[3] A. van den 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.

[4] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[5] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: a discussion and an evaluation," in *Proc. ICPhS*, 2019, pp. 487–491.

[6] K. Sjölander, J. Beskow, J. Gustafson, E. Lewin, R. Carlson, and B. Granström, "Web-based educational tools for speech technology," in *Proc. ICSLP*, 1998.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5–6, pp. 453–467, 1990.

[8] S. J. Winters and D. B. Pisoni, "Perception and comprehension of synthetic speech," *Research on Spoken Language Processing Progress Report*, no. 26, pp. 95–138, 2004.

[9] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, 2019, pp. 3617–3621.

[10] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. NeurIPS*, vol. 33, 2020.

[11] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*, 2021, pp. 6588–6592.

[12] A. Vioni, M. Christidou, N. Ellinas, G. Vamvoukakis, P. Kakoulidis, T. Kim, J. S. Sung, H. Park, A. Chalamandaris, and P. Tsiakoulis, "Prosodic clustering for phoneme-level prosody control in end-to-end speech synthesis," in *Proc. ICASSP*, 2021, pp. 5719–5723.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[14] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[15] K. Ito and L. Johnson, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.