Wavebender: A speech synthesis tool for phonetic experimentation

Gustavo Teodoro Döhler Beck, Ulme Wennberg, Gustav Eje Henter, Zofia Malisz
KTH Royal Institute of Technology, Stockholm

Synthetic speech is approaching a signal quality that is virtually indistinguishable from human speech (Malisz et al. 2019). This has come to pass by replacing signal processing with deep learning, in effect ceding control over the generated speech to the machine. Human beings, however, exhibit great control over their speech. Therefore, speech technologists have shifted research focus towards regaining control of the output speech – to improve its applicability and appropriateness in the communicative context.

A particularly interesting application of controllable synthesis is the speech sciences, in which synthetic speech has long been an important tool for stimulus creation. However, such applications require highly accurate control over the output. Modern, neural speech synthesis generally does not provide such control. As a result, speech scientists rely on legacy tools e.g., formant synthesisers. Unfortunately, such speech also sounds artificial and is processed and perceived differently from natural speech by humans. This casts doubt on the universality of research findings derived from such stimuli.

Our belief (Malisz et al. 2019) is that controllable speech synthesis for the phonetic sciences is a compelling research problem for speech technology and machine learning. We present a proof-of-concept method (Beck et al. 2022) that tries to marry the high quality of neural speech synthesis with phonetically relevant control. We train a small network with CNNs and adversarial losses to generate acoustic features from perceptually relevant speech parameters. One can then leverage a pre-trained neural vocoder to convert these acoustics into audio. The system, Wavebender GAN, can be trained to create acoustic features from any set of control parameters, but we specifically demonstrate its use by creating the first (to our knowledge) neural formant synthesiser, controlled by a minimalist parameter set:

- fundamental frequency (linearly interpolated log f0 and
- a binary voicing flag)
- formant frequencies (F1 and F2)
- two measures of voicing quality (spectral centroid and spectral slope)

HiFi-GAN (Kong et al. 2020) is used as the neural vocoder. Subjective and objective experiments demonstrate impressive speech quality and promising results in terms of control accuracy.

G. T. Doehler Beck, U. Wennberg, Z. Malisz, and G. E. Henter, 2022, *"Wavebender GAN: An architecture for phonetically meaningful speech manipulation"*, in Proc. ICASSP 2022.
Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, 2019, *"Modern speech synthesis for phonetic sciences: a discussion and an evaluation"*, in Proc. ICPhS 2019.
J. Kong, J. Kim, and J. Bae, 2020, *"HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis"*, in Proc. NeurIPS 2020.