Deep learning for phonetically meaningful speech manipulation

Gustavo Teodoro Döhler Beck[1], Ulme Wennberg[1], Zofia Malisz[1], Gustav Eje Henter[1]
[1] Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
[gtdb, ulme, malisz, ghe]@kth.se

Abstract

The quality of synthetic speech has advanced rapidly in the last decade. Unfortunately, the new technologies have rarely proven to be useful for the speech sciences community. The modern methods lack direct and accurate control over important speech properties such as formants - necessary for stimulus creation in the speech sciences. Consequently, stimulus creation currently still relies on legacy methods that are typically based on task-specific signal processing. Consequently, using manipulated stimuli with audible signal processing artefacts may result in research findings that will not generalise to human perception of natural, artefact-free speech.

This paper presents a recent system, Wavebender GAN, for manipulating phonetically meaningful speech properties via deep learning rather than custom-designed signal processing. The system learns to manipulate arbitrary acoustic properties by training on speech data with property annotations. As the system uses neural vocoders, advances in vocoder technology will automatically result in a more realistic output.

As a demonstration, we train an example system that essentially mimics a (deep learning-based) formant synthesiser. We present objective and subjective experiments that confirm the potential of our approach. We hope this work is a step towards advanced modern tools for phoneticians and strengthens the dialogue between speech science and technology.

Introduction

Our best scientific models of human speech and speech perception are a product of a fruitful dialogue between speech scientists and engineers (Malisz et al., 2019). On the one hand, speech science helped synthesis get started (King, 2014), in that early formant synthesisers relied on established models of speech production and perception (Fant, 1960). Some aspects of this perception-based approach to modelling, such as the mel scale, remain widely used also in the current era of data-driven speech technology based on machine learning. On the other hand, insights into speech sciences, for example evidence for categorical speech perception and speech perception theory, were arrived at through listening tests on synthetic speech stimuli (Liberman & Mattingly, 1985).

In recent decades, however, the two fields have taken different paths. Speech technology has focused on synthesising speech of the highest quality and realism, both with regards to speech waveform generation, e.g., van den Oord et al. (1996), Tamamori et al. (2017), Kong et al. (2020), and in text-to-speech (TTS), e.g., Shen et al. (2018). This, however, has come at the expense of controllability, by ceding control over the synthesis process and fundamental properties of the generated speech to a machine-learning algorithm, instead of leveraging knowledge-based acoustic models. In particular, modern neural vocoders and TTS systems do not offer precise, frame-level control of important acoustic cues like pitch, formant frequencies, voice/phonation quality, etc.

Many perceptual experiments in speech sciences require careful control over the above signal properties. Because modern speech technology does not offer such control, experiments need to rely on older techniques such as formant synthesis (Fant, 1960), e.g., using the system of Sjölander et al. (1998), or acoustic manipulation methods like PSOLA (Moulines & Charpentier, 1990) that do provide it. These necessary choices come with issues, e.g.: an inferior perceptual similarity to natural human speech, compared to modern speech technology. There is a comprehensive body of research cataloguing important differences in how humans perceive natural speech recordings versus how they perceive synthetic speech from legacy tools (Winters & Pisoni, 2004). These results cast doubt on the generality of research findings deduced from experiments that have used audibly unnatural-sounding speech stimuli from such tools.

This paper presents Wavebender GAN (Döhler Beck et al., 2022), an approach to controllable speech synthesis and manipulation of phonetically relevant speech properties that is based on deep learning rather than conventional signal processing. The goal is to combine the affordances for precise control that speech scientists require with the exceptional realism offered by modern speech technology and obtain the best of both worlds. For our experiments, we demonstrate a proof-of-concept system that performs formant synthesis using neural nets.

Related work

The most important puzzle piece for the approach we describe is the rise of neural vocoders, e.g., Tamamori et al. (2017), Kong et al. (2020), which are deep-learning models trained to synthesise natural-sounding speech waveforms when given a mel-spectrogram as input. Good neural vocoders often require great computational resources to train, but fortunately there are many pre-trained neural vocoders available off the shelf. A particularly convincing one is HiFi-GAN (Kong et al., 2020), which we leverage for this work.

It is possible to train neural vocoders to recreate realistic waveforms from other speech representations than mel-spectrograms. Most relevant to Wavebender GAN is perhaps the work of Juvela et al. (2018), who reconstruct natural-sounding waveforms from only 20 mel-frequency cepstrum coefficients (MFCCs) per frame of speech. Our work uses a related but different approach to further push the envelope and create convincing audio from only five numbers – the values of five different, perceptually relevant speech parameters – per frame of speech. Our solution makes use of a pre-trained neural vocoder to simplify the problem, where Wavebender GAN generates mel-spectrograms from the chosen speech parameters, and the vocoder then turns these mel-spectra into an audio signal.

Method

We now give a high-level overview of Wavebender GAN and how one can create a Wavebender GAN

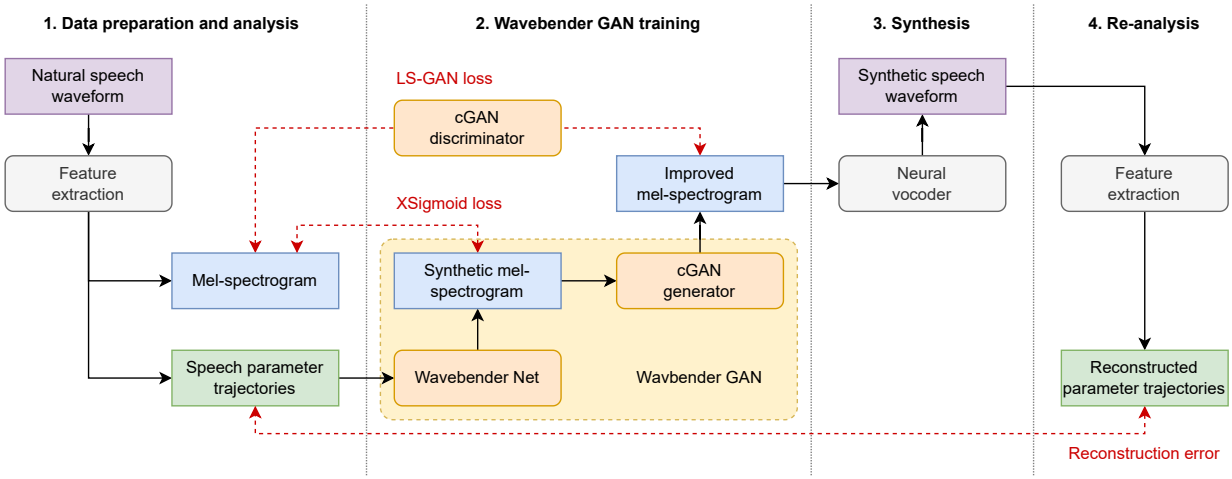

Figure 1. Workflow and pipeline for model creation and evaluation. Square boxes are data, rounded are boxes systems. Waveforms are purple, mel-spectrograms blue, and phonetically-relevant speech parameters green. Components trained in this paper are orange/yellow whilst other ones are grey. Red text and arrows denote loss functions and error measures.

system that enables speech synthesis and manipulation from an arbitrary set of speech parameters of interest. Additional technical details are provided in our main ICASSP paper (Döhler Beck et al., 2022). A graphical overview of Wavebender GAN training, synthesis, and evaluation is provided in Figure 1.

Requisite data and data preparation

Wavebender GAN is a method based on machine learning. To create a Wavebender GAN system that allows control over particular speech features (a.k.a. *speech parameters*) and that speaks in a particular voice, one needs training data for the machine to learn from. In this case, one needs studio-quality audio recordings from the relevant speaker. These recordings have to be coupled with a log-magnitude mel-spectrogram (for driving a neural vocoder) along with regular annotations of the speech parameters of interest at the same frame rate as the mel-spectrograms. These represent the outputs and the inputs of the machine learning, respectively.

Annotations can be manual, or obtained using automatic feature extractors. Although automatic feature extraction can make errors, we expect the quality of these tools to improve over time, steadily providing better data from which to build Wavebender GAN systems.

We found that *data augmentation* can improve results. This is a process of creating additional, synthetic training data (here parameter trajectories and matching mel-spectrograms) for machine learning. The synthetic data represents different parameter ranges and combinations that are underrepresented in the data or in normal speech in general. In our experiments, such augmentation was particularly important for f0, but it also improved the control over other speech properties not directly related to the specific augmentation used.

Wavebender GAN

A Wavebender GAN system consists of two subsystems, both trained using the data described above, but in different ways. Together, these produce plausible synthetic speech mel-spectrograms that express the phonetic characteristics (speech parameter trajectories) provided as input to the system. This spectrogram is then is passed to a suitable, pre-existing neural vocoder (off-the-shelf or bespoke) to create high-quality speech waveforms.

The first subsystem, called *Wavebender Net*, is a variant of a the popular ResNet architecture for image classification (He et al., 2016), but adapted to take time series of speech parameter as input and to return a mel-spectrogram as output. Initial Wavebender Net training minimises a loss function called the XSigmoid loss, which is similar to minimising the squared error, but with greater robustness against issues such as outliers.

By itself, a well-trained Wavebender Net produces mel-spectrograms that accurately express the speech parameters of interest. However, they are oversmoothed and lack the fine detail of natural speech acoustics. To improve this, we train a conditional generative adversarial network (cGAN) (Mirza & Osindero, 2014) whose generator transforms the oversmoothed mel-spectrograms from Wavebender Net into similar, but more perceptually realistic mel-spectrograms. This training uses the LS-GAN framework (Mao et al., 2017). The final Wavebender GAN system comprises the Wavebender Net and the cGAN generator put together.

Experiments

To study the abilities of the Wavebender GAN framework in speech reconstruction and manipulation, we trained an example system and performed experiments on the LJ Speech database (Ito & Johnson, 2017), containing 24 hours of text and speech in a female US English voice. 95% of the data was used for training and 5% for testing.

For the experiments, we selected a core set of five phonetically meaningful speech parameters to drive the synthesiser, specifically f0 (including voicing), F1, F2, spectral centroid, and spectral slope. These features were extracted from the data using Surfboard (Lenain et al., 2020), with Parselmouth (Jadoul et al., 2018) used for pitch manipulation as part of data augmentation. We evaluated the trained Wavebender GAN system in terms of both the control accuracy over speech properties, and in terms of subjective listener ratings of the naturalness of the synthesised speech. Evaluation details are available in Döhler Beck et al. (2022); below follows a summary of the studies and their results. Example audio stimuli are available at the project webpage at URL https://gustavo-beck.github.io/wavebender-gan/.

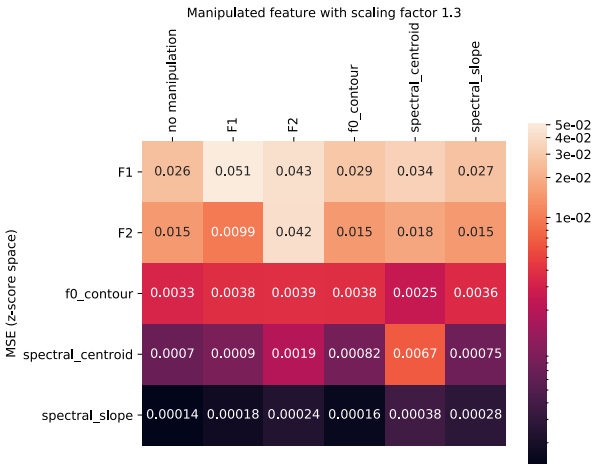

Figure 2. The effect of a factor 1.3 scaling of a single feature (columns) on the relative MSE reconstruction error of any given feature (rows). Numbers are relative MSEs; the darker the shade the smaller the error.

Accuracy of reconstruction and manipulation

In an experiment with copy synthesis, Wavebender GAN was able to reconstruct speech signals with speech parameters very similar to those provided as input (as confirmed by running Surfboard on the synthetic audio and comparing the resulting speech parameter trajectories to the input trajectories). Overall, the reconstruction errors were small. Even though Wavebender GAN uses a minimalist feature set and more networks and processing steps than HiFi-GAN, speech signals generated directly from mel-spectrograms using HiFi-GAN had very similar reconstruction errors as Wavebender GAN for all speech features. This suggests that the vocoder, despite providing good perceptual naturalness, is the main bottleneck in terms of control accuracy. By improving the vocoder, we can expect improved control, especially for the formant frequencies, which had the biggest reconstruction errors in our study, around 1 to 3 percent.

After copy synthesis, we also studied the effect of manipulating different speech features in isolation, by scaling them up or down whilst keeping all other feature tracks the same as before. When applying a moderate amount of scaling (about 0.8 to 1.2), this did not majorly affect the error in the feature trajectories from the synthetic audio, except when scaling up F2. A breakdown of the errors in different features for a scaling factor of 1.3, the most extreme case we studied, is available in Figure 2. In the figure, F1 and F2 have the greatest errors (no matter whether manipulated or not) of up to 5% in the worst case. The figure suggests successful disentanglement in most cases, since the reconstruction error of a given parameter changes little when manipulating other parameters, with the possible exception of increased F1 error when manipulating F2.

Subjective listening experiments

We performed a blinded listening test where 29 crowdsourced listeners were asked to rate the naturalness of reconstructed speech stimuli, using the classic MOS scale from 1 (bad) to 5 (excellent). (Manipulated speech was not considered here, since changes such as pitch manipulation can lead to intrinsically less natural speech audio.) A natural speech stimulus was always provided as reference. Stimuli generated from parameter trajectories using Wavebender GAN reached a mean opinion score of 4.12±0.31, not far behind the state-of-the-art HiFi-GAN neural vocoder, which scored 4.44±0.16 in the same evaluation.

Impressions from informal listening to manipulated speech stimuli suggest good pitch manipulation and that formant manipulation changes vowel perception. Manipulating the spectral centroid alters the perception of fricatives, for example producing a kind of lisp. This is consistent with prior findings from phonetics, in that spectral moments acoustically define places of articulation in English fricatives (Jongman et al., 2000).

Limitations

Wavebender is a proof-of-concept system. As such, there remain a number of unaddressed questions on the road to practical tools for everyday use in speech sciences. Answering these questions offers opportunities for collaborative research between technologists and phoneticians.

First, we have not compared the manipulation accuracy or speech realism of our Wavebender GAN system to those offered by established legacy tools such as, e.g., PSOLA (Moulines & Charpentier, 1990) for pitch manipulation, or formant synthesis (Fant, 1960) for controlling formant frequencies. Informal listening suggests our system achieves similar perceptual naturalness as existing tools on pitch-manipulation tasks. Importantly, the mean opinion scores we achieved are substantially better than what we would expect from classic formant synthesis as implemented in Praat (Boersma, 2001; Malisz et al. 2019).

To create a Wavebender GAN system from scratch for a particular application requires a sizeable database of speech where, ideally, the properties of interest have been accurately annotated. Such datasets can be expensive to create, and obtaining a large amount of data like this is not always feasible. This limitation can probably be overcome via first training a Wavebender GAN on one or more other voices with lots of data available, and then *fine-tuning* that initial model to match the speaker of interest, by continuing training only on data specifically from that speaker. This is a widely used deep-learning method for getting better results on small datasets. We hope to try this in the future, since a successful result would greatly extend the range of situations where the method can be applied. In general, the wide availability of large speech databases has enabled speech technology to generalise better to new situations without special training data, as shown by Lorenzo-Trueba et al. (2018).

Finally, we have not directly verified the utility of Wavebender GAN for speech-sciences research. One way to do so would be to use Wavebender GAN to craft stimuli for a classic perceptual experiment such as categorical perception (Van Hessen et al. 1999) and then verify that reasonable conclusions result from that experiment. Such validation is also future work.

Conclusion

We have argued that modern speech technology has overlooked the needs of speech scientists, and that recent advances in deep learning could be harnessed to provide better tools for audio stimulus creation in speech sciences. This would enable greater perceptual similarity to

natural speech and a broader palette of phonetically and perceptually interesting speech-signal properties to control, without requiring bespoke signal processing approaches and reducing manual work.

We have furthermore described a proof-of-concept system called Wavebender GAN that illustrates how these unmet needs can be addressed. Wavebender GAN is designed to be able to manipulate arbitrary speech-signal properties whilst still creating synthetic stimuli that sound like natural human speech. The system is based on deep learning, trained on a dataset of speech recordings and their corresponding speech-feature values of interest. We specifically study formant synthesis using neural vocoders as an example of the approach. Though several limitations remain to be addressed before a general-purpose tool is obtained, our empirical results are encouraging. Future advances in feature extraction and neural vocoders can only strengthen the approach. Our findings thus stake out a direction toward better speech technology for speech scientists, and to a reinvigorated dialogue between the two communities.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Zofia Malisz was supported by the Swedish Research Council grant no. 2017-02861 “Multimodal encoding of prosodic prominence in voiced and whispered speech”. The authors thank Jonas Beskow for helpful feedback.

References

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10), pp. 341–345.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague, Netherlands: Mouton.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings CVPR 2016* (pp. 770–778). Las Vegas, NV, USA. doi: 10.1109/CVPR.2016.90

A. J. Van Hessen et al. (1999). Categorical perception as a function of stimulus quality. *Phonetica*, vol. 56, no. 12, pp. 56–72.

Ito, K., & Johnson, L. (2017). *The LJ Speech dataset*. https://keithito.com/LJ-Speech-Dataset/

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71, pp. 1–15. doi: 10.1016/j.wocn.2018.07.001

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3), pp. 1252–1263. doi: doi.org/10.1121/1.1288413

Juvela, L., Bollepalli, B., Wang, X., Kameoka, H., Airaksinen, M., Yamagishi, J., & Alku, P. (2018). Speech waveform synthesis from MFCC sequences with generative adversarial networks. In *Proceedings ICASSP 2018* (pp. 5679-5683). Calgary, AB, Canada. doi: 10.1109/ICASSP.2018.8461852

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, 1(1), e006. doi: 10.3989/loquens.2014.006

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings NeurIPS 2020* (pp. 17022–17033).

Lenain, R., Weston, J., Shivkumar, A., & Fristed, E. (2020). Surfboard: Audio feature extraction for modern machine learning. *arXiv preprint arXiv:2005.08848*.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36. doi: 10.1016/0010-0277(85)90021-6

Lorenzo-Trueba, J., Drugman, T., Latorre, J., Merritt, T., Putrycz, B., Barra-Chicote, R., ... & Aggarwal, V. (2018). Towards achieving robust universal neural vocoding. *arXiv preprint arXiv:1811.06292*.

Malisz, Z., Henter, G. E., Valentini-Botinhao, C., Watts, O., Beskow, J., & Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In *Proceedings ICPhS 2019* (pp. 487–491). Melbourne, Australia.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings CVPR 2017* (pp. 2794–2802). Honolulu, HI, USA. doi: 10.1109/TPAMI.2018.2872043

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6), pp. 453–467. doi: 10.1016/0167-6393(90)90021-Z

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings ICASSP 2018* (pp. 4779-4783). Calgary, AB, Canada. doi: 10.1109/ICASSP.2018.8461368

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., & Granström, B. (1998) Web-based educational tools for speech technology. In *Proceedings ICSLP 1998*, paper 0361. Sydney, Australia.

Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K., & Toda, T. (2017). Speaker-dependent WaveNet vocoder. In *Proceedings Interspeech 2017* (pp. 1118-1122). Stockholm, Sweden. doi: 10.21437/Interspeech.2017-314

Winters, S. J., & Pisoni, D. B. (2004). Perception and comprehension of synthetic speech. *Research on Spoken Language Processing Progress Report*, 26, pp. 95–138.