WAVEBENDER GAN:
AN ARCHITECTURE FOR PHONETICALLY MEANINGFUL SPEECH MANIPULATION

Gustavo Teodoro Döhler Beck, Ulme Wennberg, Zofia Malisz, Gustav Eje Henter

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

ABSTRACT

Deep learning has revolutionised synthetic speech quality. However, it has thus far delivered little value to the speech science community. The new methods do not meet the controllability demands that practitioners in this area require e.g.: in listening tests with manipulated speech stimuli. Instead, control of different speech properties in such stimuli is achieved by using legacy signal-processing methods. This limits the range, accuracy, and speech quality of the manipulations. Also, audible artefacts have a negative impact on the methodological validity of results in speech perception studies.

This work introduces a system capable of manipulating speech properties through learning rather than design. The architecture learns to control arbitrary speech properties and leverages progress in neural vocoders to obtain realistic output. Experiments with copy synthesis and manipulation of a small set of core speech features (pitch, formants, and voice quality measures) illustrate the promise of the approach for producing speech stimuli that have accurate control and high perceptual quality.

Index Terms— speech synthesis control, speech science methods, neural vocoder, adversarial learning, speech perception

1. INTRODUCTION

Speech synthesis based on data-driven machine learning, e.g., Wave-Net [1], has been shown to exceed previous realism standards with a margin [2]. These innovations, however, have not yet been taken up into the toolbox of methods employed in basic speech and language research. The reason is that modern synthesis systems do not offer adequate control over low-level acoustic parameters, such as duration, formant frequencies, or fundamental frequency.

Phonetic and psycholinguistic experiments that aim to disentangle the perceptual role of particular acoustic features in the speech signal often involve the creation of well-controlled audio stimuli using speech synthesis. To this day, these stimuli are created by legacy systems, e.g.: formant synthesis (e.g., the Klatt synthesiser [3] or OVE [4]) or overlap-add techniques, e.g., [5]. However, speech generated by legacy systems is processed differently by humans than natural speech. It overburdens attention and cognitive mechanisms, evident through slower processing times [6]. The signal does not preserve the natural variability or acoustic-phonetic cue redundancies that are fundamental in human perception of natural speech. This casts doubt on the universality of research findings associated with using synthetic stimuli generated by legacy systems [6, 7].

In this project, we aim to create a synthesis method that combines the precise control over phonetically interpretable acoustic parameters, such as duration, formant frequencies, or pitch, provided by signal-processing methods, with the perceptual closeness to natural human speech offered by modern deep-learning techniques.

We introduce a *general* architecture designed to manipulate perceptually relevant features in a speech-in-speech-out system that leverages a pre-trained neural vocoder. This means the model we propose is not contingent on manipulating one, or a specific set, of features. In practice, we note, model performance will depend on what feature set is used, as it is important for features to be both a) phonetically meaningful and b) sufficiently disentangled to achieve speech reconstruction and manipulation, respectively. We *validate* the approach in terms signal reconstruction quality and manipulation effectiveness on a specific set of core features from phonetics. Audio examples are available at gustavo-beck.github.io/wavebender-gan/.

2. PRIOR WORK

Speech manipulation has a long history as part of the bread and butter of phonetic research. Classic formant synthesis methods [3, 4] can produce high-quality, manipulated synthetic speech [8]. However, the process is laborious and most stimuli sound distinctly unnatural [6, 9]. To manipulate pitch and duration, many signal-processing-based techniques exist, e.g.: pitch-synchronous overlap-add (PSOLA) [5]. These operate in the waveform-domain and offer good quality with noticeable perceptual artefacts. Signal processing-based vocoders such as STRAIGHT [10] allow explicit pitch manipulation and can be used in speech perception research involving voice morphing [11]. However, the manipulation of other phonetically meaningful properties is less well-served by the signal processing. For any given perceptual property of interest, manipulation methods may be poorly developed or even non-existent, as bespoke methods require substantial signal processing research.

Neural vocoders, that generate speech waveforms using deep-learning, have recently emerged as an alternative to classical vocoders such as [10]. This enabled a step change in synthetic-speech naturalness, first demonstrated by WaveNet [1] for TTS and then adapted for neural vocoding in [12]. Unlike classical vocoders based on signal processing, the mapping from input features to output waveform in a neural vocoder is learned from data rather than manually designed. The same general machine-learning approach can therefore be used to create vocoders with many different control interfaces, and the resulting signal quality is not limited by signal-processing but by the machine-learning methodology and data quality. This article proposes a concrete deep-learning framework for controlling arbitrary speech properties validated to give both high signal quality and accurate control.

To date, many high-quality neural vocoders have been proposed, harnessing a variety of deep-learning approaches including dilated
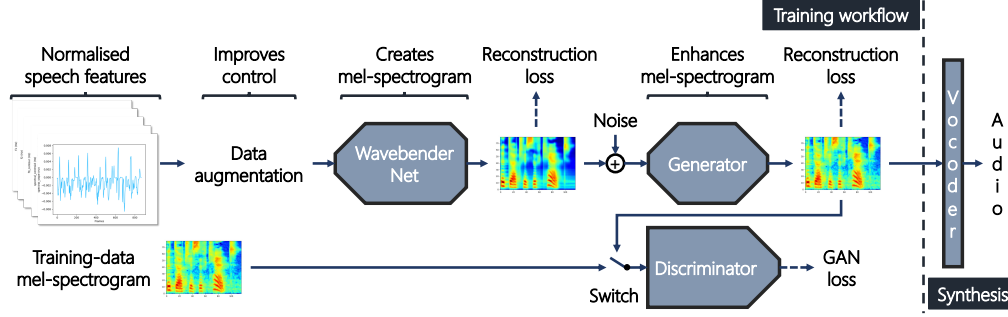
Fig. 1: Wavebender GAN workflow for training and synthesis. Network shapes in the figure are meaningful and indicate layer widths.

CNNs [1], RNNs [13, 14, 15], GANs [16, 17, 18], normalising flows [19, 20], and denoising diffusion models [21, 22]. At present, the most widely-used neural vocoder is likely WaveGlow [19], although HiFi-GAN [18] has emerged as a drop-in replacement with state-of-the-art speech signal quality.

Most neural vocoders synthesise speech waveforms from (mel) spectrogram acoustic features [23]. While these input features are easy to extract from speech and work well in text-to-speech applications, they are not an intuitive representation for speech manipulation. Neural vocoders [16, 15, 17] that leverage the source-filter model of speech production [24] offer a better starting point for pitch manipulation, but not necessarily other relevant speech properties. Our deep-learning-based approach to speech manipulation provides simultaneous control over many phonetically meaningful speech properties chosen by the experimenter, for example formant frequencies and voice-quality measures like spectral slope.

A system similar to ours is demonstrated in [25]: a GAN-based model for speaker-specific synthesis of high-quality speech waveforms, specifically, from 20 mel-frequency cepstrum coefficients (MFCCs) per analysis frame. Our work improves on [25] by a) avoiding the need to train new models at the waveform level by leveraging existing high-quality vocoders such as [18], b) achieving realistic synthesis using only 5 speech parameters per frame (plus a binary flag), thus further advancing minimalist vocoder interfaces, and c) parameterising the control in terms of phonetically relevant speech properties rather than MFCCs.

3. METHOD

We describe the proposed methodology for controllable speech synthesis from perceptually-meaningful speech parameters, which comprises two major steps:

1. A modified ResNet that generates vocoder features (here mel-spectrograms) from selected speech parameters (Sec. 3.2)
2. A post-processing network that enhances the output from the previous network, trained using a GAN loss (Sec. 3.3)

The final vocoder features (mel-spectrograms) are then fed into a pre-trained neural vocoder to generate natural-sounding synthetic speech waveforms. The proposed approach is visualised in Fig. 1.

3.1. Problem formulation

Conceptually, our task of synthesising speech from a set of (phonetically relevant) speech parameters can be cast as the inverse of feature extraction. We assume that feature extractors exist to compute all the relevant features from recorded speech, at a particular framerate. We apply the feature extractors to a speech dataset to obtain a training set that includes waveforms and associated sequences of speech parameters. While the performance of our approach is limited by the quality of the feature extractors, improved speech parameter extraction is out of scope of this work. In general, we expect the accuracy of off-the-shelf feature extractors to keep increasing, bringing steady improvements to our method.

Formally speaking, the task we consider in this work is to take a series of input vectors, representing the speech properties of interest at each time frame, and use these to generate a synthetic speech waveform. We leverage a (pre-trained neural) vocoder for waveform generation to make use of existing advances in speech signal generation. This way we reduce the original problem to the mapping of input vectors to an output vectors series (mel-spectrograms) that, in turn, constitute input features to the vocoder. From now on, we assume that the vocoder features are log-energy mel-spectrograms since these are widely used [23] and easy to extract from audio. The resulting setup amounts to a sequence-to-sequence regression problem where the input and output sequences are already time-aligned.

3.2. Wavebender Net for initial mel-spectrogram creation

The fact that inputs and output series are aligned in time suggests using a convolutional architecture to map between the two. We perform the initial mapping from input to output sequence using an architecture inspired by the successful ResNet architecture [26] from computer vision, which we adapt and enhance for the task at hand. Our modifications are 1) changing the architecture from 2D to 1D convolutions, since we are working on time series (one dimension) instead of images that have both width and height; 2) adding long-range skip connections which were found to improve learning in U-Net [27] and FlowNet [28]; and 3) replacing batchnorm by group normalisation [29] which performs better on small batches as considered here. We call this component *Wavebender Net*. Like all other networks in the paper, we make this network wider in its intermediate layers, to not limit its capacity and benefit from the lottery ticket hypothesis [30] in learning accurate mappings.

We train our model using a regression loss (the original ResNet was trained as a classifier). Specifically, we chose the XSigmoid loss function, instead of the more common L2/MSE loss, since XSigmoid has better statistical robustness to outliers (similar to [31]) and leads to sharper, less over-smoothed output. That said, we still find minor artefacts in the generated mel-spectrograms (see Fig. 4), which is why we introduce the enhancement network in the next section.

In initial experiments we noticed that Wavebender Net worked well for copy-synthesis, but that changing some of our chosen speech parameters, most noticeably pitch, away from the values in the copy synthesis reduced speech quality. It also did not accurately reflect the desired manipulation value. We hypothesised that this issue was due to a lack of representative data (e.g., high and low pitches) in

6188

the training set. We solved the problem using data augmentation: by randomly manipulating input speech and re-extracting the associated speech parameters, training examples can be created that cover a wider region of the speech parameter space. Importantly, the augmentations used need not be related to the speech parameter set used – our experiments applied utterance-level f0 manipulation and gain modification, even though gain has no effect on our speech parameters. This teaches the network to be independent of gain. Informal testing confirmed that data augmentation noticeably improved control accuracy.

3.3. Conditional GAN for enhancing mel-spectrograms

Deterministic predictors of stochastic signals, such as speech, generally exhibit averaging artefacts such as oversmoothing. To reinsert perceptual cues lost in the averaging and remove other artefacts of deterministic prediction, we use a second network (the GAN *generator*) that maps from mel-spectra to enhanced mel-spectra, based on conditional GAN (cGAN) techniques [32]. Most cGANs also take random noise as input, to enable stochastic synthesis. We accomplish this by adding a small amount of noise to each spectral bin. The result is passed through the rest of the generator, which is composed of 2D CNN layers. The discriminator also consists of 2D CNNs. We call our complete setup *Wavebender GAN*.

We use an LS-GAN loss [33] to train the generator and the discriminator. We also compute the XSigmoid reconstruction loss of the generated spectrograms before and after the generator. The final objective function is the sum of these two terms and the LS-GAN loss. The discriminator, the generator, and Wavebender Net are then trained simultaneously, end-to-end. To accelerate convergence, training is initialised from a pre-trained Wavebender Net.

4. EXPERIMENTS

4.1. Data and setup

We performed our experiments on the LJ Speech dataset [34] which comprises 13,100 utterances (approximately 24 h) from a single female speaker of US English reading books aloud. The data was randomly split into a 95% training and 5% testing utterance sets.

For the experiments reported here, we considered a core set of phonetically meaningful, frame-wise speech parameters that we identified in the features provided by the Surfboard library [35]. To construct this set, we first manually identified a subset of the Surfboard features believed to be most widely relevant to phonetic research, and then analysed their correlations using Spearman's ρ. Features were then excluded from the set guided by high numbers in the correlation matrix, since input features with high correlations was found to negatively affect manipulation robustness.

By doing so, we arrived at five phonetically meaningful features, namely: F1 (Hz), F2 (Hz), f0 (voiced/unvoiced binary flag and log frequency, linearly interpolated in unvoiced regions), spectral centroid, and spectral slope. These were used in all experiments. The formant frequencies did correlate substantially, but retaining both was nonetheless necessary to be able to perform manipulations over the entire vowel space. The features were z-score normalised (i.e., mean 0, st. dev. 1) for the network training, which also means that MSE reconstruction errors were comparable between features. Data augmentation during training was performed using Parselmouth [36] for pitch manipulation and Surfboard for speech gain.

For the networks, we used 8 residual blocks in Wavebender Net, 6 CNN layers in the generator, and 12 CNN layers in the discriminator (with linear output as proposed in LS-GAN [33]). Long-range residual connections inspired by [27, 28] were inserted into the Wavebender Net from after block 1 to before block 8 and after block 2 to before block 7. The Wavebender Net was initially trained for 20 epochs, followed by 10 epochs training of the full system. Training used Adam [37] with a cosine learning-rate scheduler (cf. [38]).

In preliminary copy-synthesis experiments, we found HiFi-GAN [18] to provide significantly better reconstruction accuracy for our selected speech features than WaveGlow [19] did, in addition to providing higher perceived signal quality, as reported in [18]. We therefore used HiFi-GAN for all experiments reported in this paper.

4.2. Evaluation

We evaluated the ability of the system to reconstruct low-level speech features, and to robustly and precisely manipulate these speech features, while simultaneously producing high-quality synthesis. Additional information such as synthesis samples is provided on the paper webpage at gustavo-beck.github.io/wavebender-gan/.

First, we performed copy-synthesis and measured speech parameter reconstruction accuracy. We used Wavebender GAN to synthesise on the basis of low-level features obtained from a natural speech recording. We then extracted the same parameters from the reconstructed speech and compared the analysis-synthesis error for each parameter. We measured the error in each z-score normalised feature for every time frame, and then averaged this metric across all time frames and samples. Results are shown in Fig. 2.

Second, we extended the copy-synthesis experiment to evaluate the effect of manipulating individual speech parameters. We did this by globally scaling each speech parameter with a scaling factor $m \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$, one parameter at a time, while keeping the remaining parameters intact (except when manipulating F1, F2 where we corrected for correlations as explained below). These scaled feature trajectories were fed into Wavebender GAN, which fed into HiFi-GAN to synthesise manipulated speech. We then computed the MSE of the desired vs. the realised feature trajectories across all normalised features, to evaluate how well the manipulated speech respected the control.

Results are illustrated in Fig. 3. In addition to that figure, which reports the error in aggregate, the paper webpage presents a matrix that quantifies the effect of these manipulations on each individual speech parameter, as a way to study parameter disentanglement.

As explained in Sec. 4.1, F1 and F2 are correlated and may never be decoupled completely (e.g., F2 < F1 is impossible by definition). Several approaches to handling this are possible, but here (including Fig. 3) we allowed the user to manipulate either one of the features (e.g., the F1), and then use a neural-network architecture similar to Wavebender Net to predict the induced manipulation on the other feature (i.e., a corresponding F2 trajectory that respects the correlation between the two features) to arrive at the final feature trajectories for the manipulation experiment.

Third, we complemented our objective evaluation of synthesis accuracy, with and without signal manipulation, with a subjective listening test. We selected 10 natural speech clips at random from the test set. We then performed a Mean Opinion Score (MOS) test, comparing the listener preferences for speech generated by HiFi-GAN copy-synthesis and Wavebender GAN copy-synthesis (with HiFi-GAN as the vocoder), relative to the natural speech recording as a reference. In each trial, listeners were presented with the reference first. Next, they were asked to rate one of the system-generated signals presented consecutively in a counterbalanced order (N=20
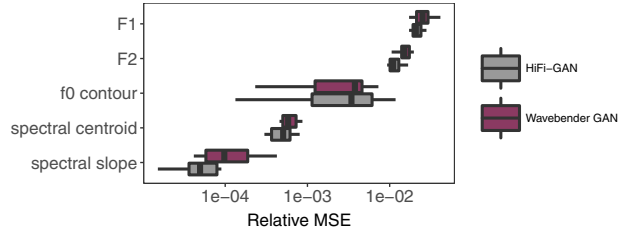
Fig. 2: Copy-synthesis error for each speech property for Wavebender GAN→HiFi-GAN vocoder versus HiFi-GAN only.
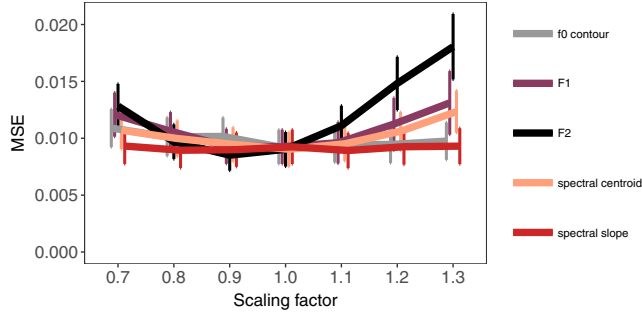

Fig. 3: Overall MSE (mean and standard deviation intervals) away from the desired value when using Wavebender GAN to manipulate (scale) each speech feature in isolation.

ratings per participant). We did not specify which system generated a given stimulus. The listeners rated signal naturalness on a 5 point scale (1-bad, 2-poor, 3-fair, 4-good, 5-excellent). Twenty nine listeners completed the survey set up on the Phonic crowdsourcing platform (N=580 total ratings). The results are reported in Sec. 5.3.

5. RESULTS AND DISCUSSION

5.1. Objective evaluation of reconstruction accuracy

Fig. 4 compares example mel-spectrograms, in a copy-synthesis setting, produced by Wavebender Net (4a) and by the entire Wavebender GAN pipeline (4b) to the mel-spectrogram of target natural speech (4c). We can see Wavebender GAN successfully removes subtle artefacts present in the Wavebender Net output.

Moreover, the results presented in Fig. 2 indicate that Wavebender GAN adds a relatively minor amount to the reconstruction errors of the different features. All errors in this paper are measured on the normalised features, to make reported values more comparable across features. The majority of the error is attributable to the vocoder (HiFi-GAN) used in the current experiments, suggesting that the vocoder, not Wavebender GAN, could be the main factor limiting control accuracy. Also, it is apparent that the data augmentation is effective for improving quality, since the Wavebender GAN error for f0 contours is smaller relative to HiFi-GAN.

5.2. Objective evaluation of manipulation effectiveness

Fig. 3 shows how the overall error across all features changes as different individual features are manipulated away from pure copy synthesis to perform speech manipulation. We see that changes to voice-quality features or the f0 contour are well reflected in the output speech and do not inflate the overall error. Manipulations of F1
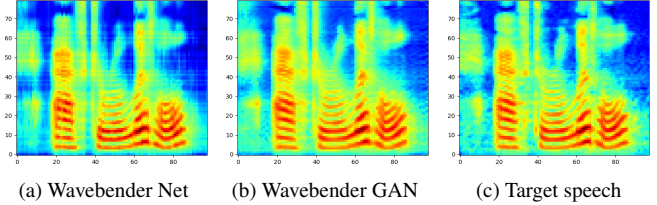

(a) Wavebender Net (b) Wavebender GAN (c) Target speech

Fig. 4: Mel-spectrogram representations generated by Wavebender Net and Wavebender GAN compared to the target representation.

and F2, however, are not as accurately reflected in the output features. We attribute this to the lack of feature augmentation for formant frequencies, in contrast to f0. A simple alternative to explicit augmentation for improving the coverage of the formant space (and the input space in general) would be to consider a large database of speakers of different genders, accents, and even languages, similar to the development of so-called universal neural vocoders [39].

5.3. Subjective MOS listening test

First of all, the mean opinion score for the HiFi-GAN copy-synthesis pipeline approximately matches the performance reported in [18] (4.36 ± 0.07), despite differences in methodology. The Wavebender GAN→HiFi-GAN vocoder copy-synthesis pipeline achieves a mean 4.12 ± 0.31 rating in the MOS test, which does not substantially deviate from the mean rating obtained by the signals produced by HiFi-GAN copy-synthesis (4.44 ± 0.16), and exceeds WaveGlow (3.81 ± 0.08) in [18]. In addition to speech parameter control, these results mean that we obtained the high output naturalness – close to a state-of-the-art neural vocoder – that we had set out achieve.

6. CONCLUSIONS AND FUTURE WORK

We have demonstrated a deep-learning approach to synthesising high-quality speech from a small set of phonetically meaningful speech parameters, as well as the ability to accurately control these parameters in the output speech. To the best of our knowledge, our pipeline is the first architecture to combine state-of-the-art speech quality with this interpretability and low-level feature control. Together, these two properties are what is needed to realise a step change in the ability of speech scientists to construct end-to-end pipelines for stimulus creation and testing of phonological models.

The main performance limitation appears not to sit in our architecture, but in the neural vocoder. As neural vocoder technology is advancing rapidly at present, we expect that further gains in quality and accuracy will follow without altering our fundamental approach.

We see several directions for future research, beyond changing the general machine-learning approach, we plan to: i) validate the model on more speakers and other feature sets such as ones that would allow to manipulate voice modes e.g.: whispered or Lombard speech, ii) extend the approach to multiple speakers, both to broaden applicability and to improve control by training on data that better covers all of speech, and iii) demonstrate the utility of the approach in performing phonetic research.

7. REFERENCES

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, et al., "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4799–4783.

[3] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[4] R. Carlson, B. Granström, and S. Hunnicutt, "A multi-language text-to-speech module," in *Proc. ICASSP*, 1982, vol. 7.

[5] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proc. ICASSP*, 1986, vol. 11, pp. 2015–2018.

[6] 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.

[7] A. J. Van Hessen and M. E. H. Schouten, "Categorical perception as a function of stimulus quality," *Phonetica*, vol. 56, no. 1-2, pp. 56–72, 1999.

[8] M. J. Hunt, D. A. Zwierzyriski, and R. C. Can, "Issues in high quality LPC analysis and synthesis," in *Proc. Eurospeech*, 1989, pp. 2348–2351.

[9] 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.

[10] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.

[11] H. Kawahara, "Temporally variable multi attribute morphing of arbitrarily many voices for exploratory research of speech prosody," in *Speech prosody in speech synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, pp. 109–120. Springer, 2015.

[12] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[13] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, et al., "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, 2017.

[14] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, et al., "Efficient neural audio synthesis," in *Proc. ICML*, 2018, pp. 2410–2419.

[15] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.

[16] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. Interspeech*, 2017.

[17] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram," in *Proc. Interspeech*, 2019, pp. 694–698.

[18] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 17022–17033.

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

[20] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet: A generative flow for raw audio," in *Proc. ICML*, 2019.

[21] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, 2021.

[22] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, et al., "WaveGrad: Estimating gradients for waveform generation," in *Proc. ICLR*, 2021.

[23] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[24] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[25] L. Juvela, B. Bollepalli, X. Wang, H. Kameoka, M. Airaksinen, et al., "Speech waveform synthesis from MFCC sequences with generative adversarial networks," in *Proc. ICASSP*, 2018.

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

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[28] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, et al., "FlowNet: Learning optical flow with convolutional networks," in *Proc. CVPR*, 2015, pp. 2758–2766.

[29] Y. Wu and K. He, "Group normalization," in *Proc. ECCV*, 2018, pp. 3–19.

[30] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *Proc. ICLR*, 2019.

[31] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, et al., "Robust TTS duration modelling using DNNs," in *Proc. ICASSP*, 2016, pp. 5130–5134.

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

[33] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. CVPR*, 2017, pp. 2794–2802.

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

[35] R. Lenain, J. Weston, A. Shivkumar, and E. Fristed, "Surfboard: Audio feature extraction for modern machine learning," *arXiv preprint arXiv:2005.08848*, 2020.

[36] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phonetics*, vol. 71, pp. 1–15, 2018.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[38] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. NeurIPS*, 2020.

[39] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, et al., "Towards achieving robust universal neural vocoding," *arXiv preprint arXiv:1811.06292*, 2018.