Adversarial Audio Synthesis with Complex-valued Polynomial Networks

Yongtao Wu [†] **Grigorios G Chrysos** [†] **Volkan Cevher** [†]

Abstract

Time-frequency (TF) representations in audio synthesis have been increasingly modeled with real-valued networks. However, overlooking the complex-valued nature of TF representations can result in suboptimal performance and require additional modules (e.g., for modeling the phase). To this end, we introduce complex-valued polynomial networks, called APOLLO, that integrate such complex-valued representations in a natural way. Concretely, APOLLO captures high-order correlations of the input elements using high-order tensors as scaling parameters. By leveraging standard tensor decompositions, we derive different architectures and enable modeling richer correlations. We outline such architectures and showcase their performance in audio generation across four benchmarks. As a highlight, APOLLO results in 17.5% improvement over adversarial methods and 8.2% over the state-of-the-art diffusion models on SC09 dataset in audio generation. Our models can encourage the systematic design of other efficient architectures on complex field.

1. Introduction

Generative Adversarial Networks (GANs) enable photo-realistic synthesis in image-related tasks [15; 27; 31; 2; 6]. Their stellar performance has prompted their use in unconditional audio synthesis, which aims to synthesize consistent utterances from noise [9; 11; 39; 45]. However, the human perception is sensitive to both global and local coherence of the waveform [11], which makes audio synthesis an inherently challenging task. We argue that the design choices, i.e., the audio representations and the network architecture, hold a key role in successful audio synthesis.

Raw waveform is primarily used for unconditional speech generation [9] while most recent studies focus on TF representation due to its theoretical expressivity and increased performance [11; 39; 45; 21]. In the TF representation, raw waveform is usually transformed through the Short-time Fourier transform (STFT) to frequency domain, which is expressed with complex numbers. To avoid using complex numbers often the phase information is discarded and only the magnitude is maintained in networks [39; 21; 45], which deteriorates the performance and phase coherence [11]. Importantly, without phase information, the TF representation is not invertible. This raises the question: *How can we explicitly model the complex-valued TF representation?*

Even though the complex-valued TF representation can be expressed by real-valued numbers and processed in real-valued neural network with two-channel outputs, a more natural representation is to design complex-valued neural network (CVNN). While CVNN has demonstrated higher generalization ability and richer capacity [25; 59; 3; 69], it has yet to demonstrate state-of-the-art performance on audio generation. On the other hand, recent theoretical advances [29; 12] prove that models with second-degree polynomials enlarge the set of functions that can be represented exactly with zero error. Polynomial nets have demonstrated flexibility and efficiency over standard neural networks in various tasks e.g., image generation [31; 6], image recognition [65], reinforcement learning [29], and sequence modeling [56]. This motivates us to design a class of polynomial nets, called APOLLO, that extracts complex-valued representations for audio generation. APOLLO expresses a complex-valued output as a high-degree polynomial expansion of the complex-valued input, as illustrated in Figure 1. Overall, our contributions can be summarized as follows:

- We introduce a new class of complex-valued polynomial nets and reveal how different architectures can be obtained by changing the factorization of the unknown parameters in the polynomial expansion.
- We conduct a thorough evaluation on audio generation and showcase the advantage of APOLLO when compared with the prior art.
- APOLLO is extended in case of multiple inputs and utilized in conditional generation. Additionally, we investigate the efficacy of learning shared representations on multimodal generation (image-to-speech).

Due to the restricted space, we include the related work in Appendix A.

[†]LIONS, EPFL, Switzerland. Correspondence to: Yongtao Wu <yongtao.wu@epfl.ch>.

Published at the workshop on Machine Learning for Audio Synthesis at the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA.

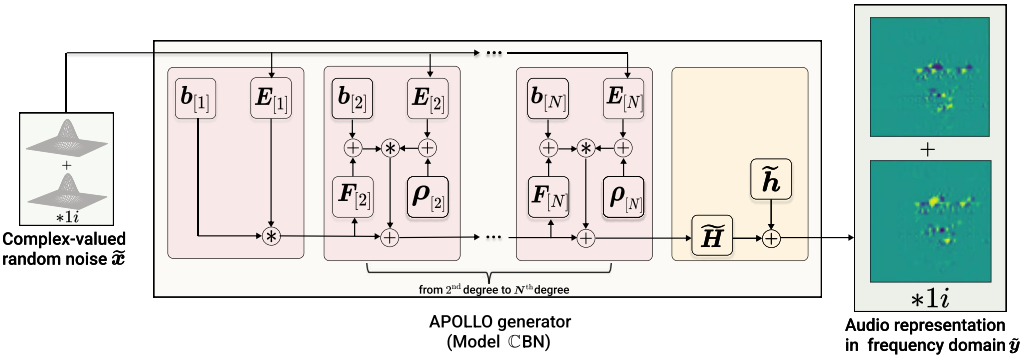

Figure 1. In this work we propose a class of functions, called APOLLO, where the complex-valued output is a polynomial of the complex-valued input. The input of the APOLLO generator is complex-valued noise and the output is the representation of audio in the frequency domain (e.g., STFT). All learnable parameters inside the pink blocks (yellow blocks) are real-valued (complex-valued).

2. Methods

Notation Real-valued vectors/matrices/tensors are symbolized by lowercase/uppercase/calligraphic boldface letters, e.g., $\boldsymbol{y}, \boldsymbol{Y}, \boldsymbol{\mathcal{Y}}$. All complex-valued variables are symbolized with wide tilde, e.g., $\widetilde{\boldsymbol{y}}$, $\widetilde{\boldsymbol{Y}}$, $\widetilde{\boldsymbol{\mathcal{Y}}}$. We denote by $*$ the Hadamard product.

2.1. Complex-valued polynomial networks

According to Mergelyan's Theorem [52], any smooth complex-valued function could be approximated by a polynomial. Our goal is to learn an N^{th} degree polynomial expansion with respect to the input $\widetilde{\boldsymbol{x}} \in \mathbb{C}^d$ with an $o-$dimensional output $\widetilde{\boldsymbol{y}}$ based on the recursive form:

$$\widetilde{\boldsymbol{y}}_n = \left(\boldsymbol{E}_{[n]}^T \widetilde{\boldsymbol{x}} + \boldsymbol{\rho}_{[n]}\right) * \left(\boldsymbol{F}_{[n]}^T \widetilde{\boldsymbol{y}}_{n-1} + \boldsymbol{b}_{[n]}\right) + \widetilde{\boldsymbol{y}}_{n-1}, \quad (1)$$

for $n = 2, \ldots, N$ with $\widetilde{\boldsymbol{y}}_1 = (\boldsymbol{E}_{[1]}^T \widetilde{\boldsymbol{x}}) * \left(\boldsymbol{b}_{[1]}\right)$ and $\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{H}} \widetilde{\boldsymbol{y}}_N + \widetilde{\boldsymbol{h}}$, where $\widetilde{\boldsymbol{H}} \in \mathbb{C}^{o \times k}$, $\widetilde{\boldsymbol{h}} \in \mathbb{C}^o$, $\boldsymbol{E}_{[n]} \in \mathbb{R}^{d \times k}$ $\boldsymbol{b}_{[n]} \in \mathbb{R}^k$, $\boldsymbol{F}_{[n]} \in \mathbb{R}^{k \times k}$, $\boldsymbol{\rho}_{[n]} \in \mathbb{R}^k$. Note that all learnable parameters in all degrees are real-valued except in the highest degree. Apart from the aforementioned models, we can also design several models based on different decomposition or the field ($\mathbb{R}$ or $\mathbb{C}$) of the parameters, we defer these models to Appendix C for completion.

2.2. Conditional complex-valued polynomial networks

The aforementioned polynomial expansions rely on a single input variable, however often there are additional variables available, e.g., class-label information. In this case, we can design polynomial expansions from multiple input variables. Motivated by the real-valued multivariate analysis [6], we focus on the case of two complex-valued inputs $\widetilde{\boldsymbol{x}} \in \mathbb{C}^d$ and $\widetilde{\boldsymbol{\psi}} \in \mathbb{C}^{d'}$. Our goal turns to learn an N^{th} degree polynomial expansion with an $o-$dimensional output $\widetilde{\boldsymbol{y}}$ with two inputs based on the following recursive relationship:

$$\widetilde{\boldsymbol{y}}_n = \left(\boldsymbol{U}_{[n,I]}^T \widetilde{\boldsymbol{x}} + \boldsymbol{U}_{[n,II]}^T \widetilde{\boldsymbol{\psi}}\right) * \widetilde{\boldsymbol{y}}_{n-1} + \widetilde{\boldsymbol{y}}_{n-1} \quad (2)$$

for $n = 2, \ldots, N$ with $\widetilde{\boldsymbol{y}}_1 = \boldsymbol{U}_{[1,I]}^T \widetilde{\boldsymbol{x}} + \boldsymbol{U}_{[1,II]}^T \widetilde{\boldsymbol{\psi}}$ and $\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{H}} \widetilde{\boldsymbol{y}}_N + \widetilde{\boldsymbol{h}}$, where $\boldsymbol{U}_{[n,I]} \in \mathbb{R}^{d \times k}, \boldsymbol{U}_{[n,II]} \in \mathbb{R}^{d' \times k}$. Note that other decompositions discussed in Appendix C can also be used.

2.3. Adversarial audio generation

In the majority of our experimental validation we use GANs, where APOLLO is chosen as the generator while the discriminator is a standard ResNet. Wasserstein loss with gradient penalty is used as the criterion of GAN due to its stability and robustness [18]. On unconditional audio generation, we implement the generator using single-variable models, e.g., Equation (1). The generator receives a complex-valued noise and outputs the representation of audio in the frequency domain, as illustrated in Figure 1. Given an audio clip, we apply STFT and truncate the Nyquist bin to obtain the complex-valued representation. In view of previous work that models the spectrum in log-scale to facilitate training, we take the square root of the absolute value of the real (and imaginary) part of the STFT and keep its sign.

When class label is available, i.e., given a noise vector $\widetilde{\boldsymbol{x}} \in \mathbb{C}^d$ and one-hot label vector $\widetilde{\boldsymbol{\psi}} \in \mathbb{C}^{d'}$ with zero imaginary part, we can implement the generator based on Equation (2), as depicted in Figure 5 of the appendix.

Previous methods focus exclusively on a single modality, i.e., audio. However, as humans we perceive information using varying sources from the real-world. To this end, we extend our APOLLO to multimodal generation (Image-to-audio). A schematic of the generator is visually depicted in Figure 2. Specifically, we first use two low-degree APOLLOs for the random noise and the image, respectively. A high-degree conditional APOLLO is utilized to capture the

Table 1. Comparison with adversarial methods on unconditional generation. Higher IS (lower FID, NDB, JSD) indicates better performance. The symbol '# par' abbreviates the number of parameters (in millions). APOLLO improves upon all the baselines in all metrics. Moreover, APOLLO-Small achieves similar performance with the baselines while reducing parameters by more than 87%.

Unconditional audio generation on SC09 dataset					
Model	IS (↑)	FID (↓)	NDB (↓)	JSD (↓)	# par (M)
Real data	8.01	0.50	0.00	0.011	-
WaveGAN [9]	4.67	41.60	16.00	0.094	36.5
SpecGAN [9]	6.03	-	-	-	36.5
TiFGAN [39]	5.97	26.70	6.00	0.051	42.4
Mel-Spec GAN [19]	5.76	-	-	-	-
BigGAN [21]	6.17	24.72	-	-	-
Π-Nets [7]	6.59	13.01	4.40	0.048	45.9
APOLLO, Small	6.48	18.90	4.20	0.038	4.6
APOLLO	**7.25**	**8.15**	**3.20**	**0.029**	64.1

correlations between the outputs of these two low-degree APOLLOs and generate the complex-valued representation of audio.

Difference from Π-Nets: APOLLO differs substantially from Π-Nets in: a) The new decomposition that yield Equation (1) is designed for reducing the number of parameters when extending to complex field and increasing the expressivity with the new bias term. b) We design architectures and technique for audio generation in Section 2.3 while Π-Nets is mostly focused on image-related tasks. c) Π-Nets have been used for a single variable input, while we also demonstrate experiments with two variables. e.g., conditional generation and multi-modal generation. Further discussion can be found in Appendix C.3.

3. Experiments

We conduct a series of experiments on audio generation to evaluate our framework from Section 3.1 to 3.4. Further details on the dataset, evaluation metrics, and experimental setup are offered in Appendix E.

3.1. Comparison against adversarial-based models

Unconditional audio generation. Firstly, we evaluate APOLLO on unconditional audio generation on three datasets used in Donahue et al. [9]: Speech Commands Zero Through Nine (SC09), Piano, Drum. The log spectrums of audios synthesized by our model are presented in Figure 3. Quantitative evaluations on SC09 dataset are reported in Table 1. APOLLO improves largely upon all the baselines in Inception Score (IS) [53], Frechet Inception Distance (FID) [24], Number of Statistically-Different Bins (NDB), and Jensen-Shannon Divergence (JSD) [50]. The corresponding 'Small' model performs similarly to the baselines while reducing parameters by more than 87%. The improvement in Piano dataset is presented in Table 2.

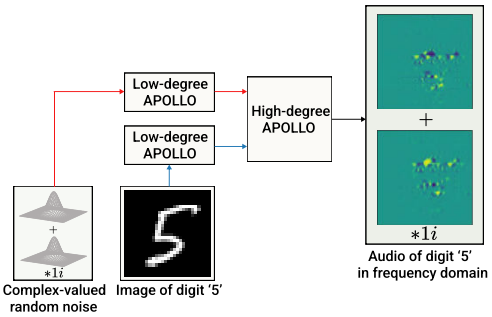

Figure 2. Generator used in image-to-speech experiments. A two-variable, high-degree APOLLO is utilized to capture the correlations of the two input variables.

Table 2. Quantitative evaluation on Piano dataset. APOLLO outperforms the compared baselines by a large margin.

Unconditional audio generation on Piano dataset			
Model	NDB (↓)	JSD (↓)	# par (M)
Real data	0.00	0.008	-
WaveGAN	24.00	0.547	36.5
TiFGAN	17.60	0.332	42.4
APOLLO, Small	13.20	0.270	**11.3**
APOLLO	**8.80**	**0.157**	42.7

Conditional audio generation. Next, we examine APOLLO in conditional audio generation on SC09 dataset and NSynth dataset [10]. The results in Table 3 demonstrate the improvement over conditional BigGAN [21] and conditional Mel-Spec GAN [19] on SC09 dataset. The result on Nsynth dataset are presented in Table 4. APOLLO improves upon GANSynth [11] in terms of FID, NDB, and JSD. It is rather remarkable that GANSynth is trained with a batch size of 8 with 11 millions samples as reported in Engel et al. [11] while APOLLO is trained with the same batch size with only 0.48 millions samples. Furthermore, APOLLO has 24% fewer parameters than GANSynth.

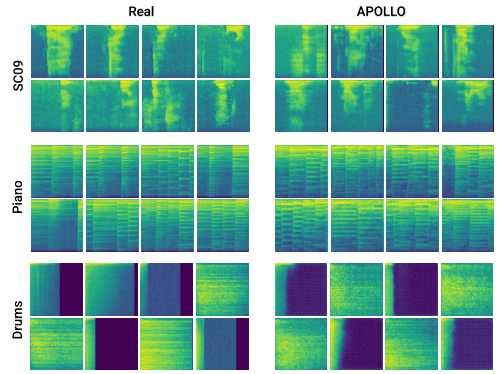

Figure 3. The log spectrums of the real samples and generated samples. For each image, the horizontal (vertical) axis is along the time (frequency). The frequency increases with interval scale from 0 HZ (top) to 8000 HZ (bottom).

3.2. Comparison against non-adversarial models

Previously, we have shown the comparison between our APOLLO and other adversarial-based models with TF representation and have exhibited the excellent performance of APOLLO. As a compensation, we compare APOLLO with non-adversarial models using waveform representation, e.g., SampleRNN [40], WaveNet [43], DiffWave [35], SASHIMI [14]. This experiment of unconditional generation is conducted on SC09 dataset. Details on the dataset splitting and evaluation metrics e.g., Modified Inception Score [20], AM Score [71], Inception Score (IS) [53], Fréchet Inception Distance (FID) [24] are the same as in Goel et al. [14]. The results in Table 7 (appendix) show that APOLLO outperforms the baselines by a large margin.

3.3. Multimodal generation: Image-to-speech

In this section, we access APOLLO in multimodal generation. We select SC09 as a source of digit audios and MNIST dataset [36] as a source of digit images. The results in Table 5 indicate that the best model achieves 72% accuracy. This experiment is more challenging than the corresponding class-conditional generation, owing to the different modality of the input-output pair, instead of the clean one-hot labels provided in the class-conditional generation. That explains the decrease in the score.

Table 3. Quantitative evaluation on conditional audio generation. APOLLO improves upon baselines by a considerable margin.

Conditional audio generation on SC09 dataset		
Model	IS (↑)	FID (↓)
Real data	8.01	0.50
BigGAN	7.33	24.40
Mel-Spec GAN	7.64	-
APOLLO	**7.73**	**6.31**

Table 4. Quantitative evaluation on NSynth. '#sam' abbreviates the total number of samples used during training (in millions). APOLLO improves upon GANSynth significantly.

Conditional audio generation on Nsynth dataset					
Model	FID(↓)	NDB (↓)	JSD (↓)	#sam (M)	# par (M)
Real data	1.44	0.00	0.002	-	-
GANSynth	3.91	30.20	0.362	11.00	14.1
APOLLO	**1.98**	**27.40**	**0.298**	**0.48**	**10.6**

Table 5. Quantitative evaluation on Image-to-speech generation on MNIST-SC09 dataset. '#acc' abbreviates the categorial accuracy.

Image-to-speech generation on MNIST-SC09 dataset			
Model	IS (↑)	FID (↓)	#acc (↑)
Real data	8.01	0.50	0.93
APOLLO, Small	5.75	26.5	0.68
APOLLO	**6.90**	**9.58**	**0.72**

3.4. Further analysis

We conduct further studies and comparisons on models trained on SC09 for unconditional generation.

Inference speed. The comparison of inference speed can be found in Appendix G.1. Even though APOLLO can directly output the STFT without phase reconstruction, APOLLO has an augmented inference time due to the complex operations, e.g., complex multiplication. A future step for our model would be to further accelerate the complex multiplications, e.g., by implementing them directly in BLAS, instead of the high-level python operations.

Human study. We invite volunteers to assign an ordinal-scale score (1 to 5) to each audio clip based on the sound quality and perception. The qualitative results are summarized in Appendix G.2. Our model obtains the highest Mean Opinion Score (MOS) [49] with respect to the prior art.

Ablation study. A thorough self-evaluation, e.g. interpolation of the inputs or empirical comparison of between different derivations, is deferred to Appendix F.

4. Conclusion

In this work, we propose APOLLO, a high degree complex-valued polynomial expansion for audio generation. To validate the architectures, we conduct thorough experiments in audio generation. APOLLO outperforms all the prior art by a large margin demonstrating the expressivity of the proposed complex-valued polynomial expansions. We believe this class of functions will be beneficial for synthesizing long audio tracks in the future. Furthermore, our experiments on conditional generation highlight the efficacy of APOLLO on multimodal generation, where the extension to large-scale models, e.g., realistic text-to-speech translation, can be an interesting application for future work.

Acknowledgements

We thank the reviewers for their thoughtful and constructive feedback. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 725594 - time-data). This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. This work was partly supported by Zeiss.

References

[1] Arjovsky, M., Shah, A., and Bengio, Y. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2016.

[2] Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[3] Choi, H., Kim, J., Huh, J., Kim, A., Ha, J., and Lee, K. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations (ICLR)*, 2019.

[4] Choi, H.-S., Kim, J., Huh, J., Kim, A., Ha, J.-W., and Lee, K. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations (ICLR)*, 2019.

[5] Chrysos, G., Moschoglou, S., Panagakis, Y., and Zafeiriou, S. Polygan: High-order polynomial generators. *arXiv preprint arXiv:1908.06571*, 2019.

[6] Chrysos, G., Georgopoulos, M., and Panagakis, Y. Conditional generation using polynomial expansions. In *Advances in neural information processing systems (NeurIPS)*, 2021.

[7] Chrysos, G. G., Moschoglou, S., Bouritsas, G., Panagakis, Y., Deng, J., and Zafeiriou, S. P-nets: Deep polynomial neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Chrysos, G. G., Moschoglou, S., Bouritsas, G., Deng, J., Panagakis, Y., and Zafeiriou, S. P. Deep polynomial neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2021.

[9] Donahue, C., McAuley, J. J., and Puckette, M. S. Adversarial audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[10] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., and Norouzi, M. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

[11] Engel, J. H., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. GANSynth: Adversarial neural audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[12] Fan, F.-L., Li, M., Wang, F., Lai, R., and Wang, G. Expressivity and trainability of quadratic networks. *arXiv preprint arXiv:2110.06081*, 2021.

[13] Giles, C. L. and Maxwell, T. Learning, invariance, and generalization in high-order neural networks. *Applied optics*, 1987.

[14] Goel, K., Gu, A., Donahue, C., and Ré, C. It's raw! audio generation with state-space models. *arXiv preprint arXiv:2202.09729*, 2022.

[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*, 2014.

[16] Griffin, D. and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.

[17] Gu, A., Goel, K., and Re, C. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

[18] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems (NeurIPS)*, 2017.

[19] Gunasekaran, D., Venkatraj, G., Brophy, E., and Ward, T. E. Improved speech synthesis using generative adversarial networks. In *AICS*, 2020.

[20] Gurumurthy, S., Kiran Sarvadevabhatla, R., and Venkatesh Babu, R. Deligan: Generative adversarial networks for diverse and limited data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 166–174, 2017.

[21] Haque, K. N., Rana, R., Hansen, J. H., and Schuller, B. Guided generative adversarial neural network for representation learning and high fidelity audio generation using fewer labelled audio data. *arXiv preprint arXiv:2003.02836*, 2020.

[22] Hardoon, D. R., Szedmak, S. R., and Shawe-taylor, J. R. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 2004.

[23] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems (NeurIPS)*, 2017.

[25] Hirose, A. and Yoshida, S. Comparison of complex- and real-valued feedforward neural networks in their generalization ability. In *International Conference on Neural Information Processing*, pp. 526–531, 2011.

[26] Hu, Y. and Loizou, P. C. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16 (1):229–238, 2007.

[27] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] Ivakhnenko, A. G. Polynomial theory of complex systems. *IEEE transactions on Systems, Man, and Cybernetics*, 1971.

[29] Jayakumar, S. M., Czarnecki, W. M., Menick, J., Schwarz, J., Rae, J., Osindero, S., Teh, Y. W., Harley, T., and Pascanu, R. Multiplicative interactions and where to find them. In *International Conference on Learning Representations (ICLR)*, 2020.

[30] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., and Kavukcuoglu, K. Efficient neural audio synthesis. In *International Conference on Machine Learning (ICML)*, 2018.

[31] Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] Kim, M. and Guest, C. Modification of backpropagation networks for complex-valued signal processing in frequency domain. In *International Joint Conference on Neural Networks (IJCNN)*, 1990.

[33] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[34] Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM review*, 2009.

[35] Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[36] LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[37] Li, C.-K. A sigma-pi-sigma neural network (spsnn). *Neural Processing Letters*, 2003.

[38] Macartney, C. and Weyde, T. Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*, 2018.

[39] Marafioti, A., Perraudin, N., Holighaus, N., and Majdak, P. Adversarial generation of time-frequency features with application in audio synthesis. In *International Conference on Machine Learning (ICML)*, 2019.

[40] Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., and Bengio, Y. Samplernn: An unconditional end-to-end neural audio generation model. 2017.

[41] Miyato, T. and Koyama, M. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018.

[42] Oh, S.-K., Pedrycz, W., and Park, B.-J. Polynomial neural networks architecture: analysis and design. *Computers & Electrical Engineering*, 2003.

[43] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[44] Oyallon, E. and Mallat, S. Deep roto-translation scattering for object classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[45] Palkama, K., Juvela, L., and Ilin, A. Conditional Spoken Digit Generation with StyleGAN. In *Interspeech*, 2020.

[46] Pruša, Z. and Søndergaard, P. L. Real-time spectrogram inversion using phase gradient heap integration. In *Proc. Int. Conf. Digital Audio Effects (DAFx-16)*, 2016.

[47] Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in neural information processing systems (NeurIPS)*. 2017.

[48] Reichert, D. P. and Serre, T. Neuronal synchrony in complex-valued deep networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[49] Ribeiro, F. P., Florêncio, D. A. F., Zhang, C., and Seltzer, M. L. Crowdmos: An approach for crowdsourcing mean opinion score studies. *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2011.

[50] Richardson, E. and Weiss, Y. On gans and gmms. Curran Associates Inc., 2018.

[51] Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2001.

[52] Rudin, W., RUDIN, W., and Company, T. M.-H. P. *Real and Complex Analysis*. McGraw-Hill Education, 1987.

[53] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. Improved techniques for training gans. In *Advances in neural information processing systems (NeurIPS)*, 2016.

[54] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pp. 4779–4783, 2018.

[55] Shin, Y. and Ghosh, J. The pi-sigma network: An efficient higher-order neural network for pattern classification and function approximation. In *International Joint Conference on Neural Networks (IJCNN)*, 1991.

[56] Su, J., Byeon, W., Kossaifi, J., Huang, F., Kautz, J., and Anandkumar, A. Convolutional tensor-train lstm for spatio-temporal learning. *Advances in neural information processing systems (NeurIPS)*, 33:13714–13726, 2020.

[57] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[58] Thiemann, J., Ito, N., and Vincent, E. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, pp. 035081, 2013.

[59] Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J. F., Mehri, S., Rostamzadeh, N., Bengio, Y., and Pal, C. J. Deep complex networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[60] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. Complex embeddings for simple link prediction. In *International Conference on Machine Learning (ICML)*, 2016.

[61] Tygert, M., Bruna, J., Chintala, S., LeCun, Y., Piantino, S., and Szlam, A. A mathematical motivation for complex-valued convolutional networks. *Neural computation*, 28(5):815–825, 2016.

[62] Valle, R., Shih, K. J., Prenger, R., and Catanzaro, B. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[63] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016.

[64] Veaux, C., Yamagishi, J., and King, S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pp. 1–4, 2013.

[65] Wang, X., Girshick, R., Gupta, A., and He, K. Non-local neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[66] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. Tacotron: Towards end-to-end speech synthesis. *INTERSPEECH*, 2017.

[67] Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[68] Xiong, Y., Wu, W., Kang, X., and Zhang, C. Training pi-sigma network by online gradient algorithm with penalty for small weight update. *Neural computation*, 2007.

[69] Yang, M., Ma, M. Q., Li, D., Tsai, Y. H., and Salakhutdinov, R. Complex transformer: A framework for modeling complex-valued sequence. In *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2020.

[70] Zhang, S., Gong, Y., and Yu, D. Encrypted speech recognition using deep polynomial networks. *CoRR*, 2019.

[71] Zhou, Z., Cai, H., Rong, S., Song, Y., Ren, K., Zhang, W., Wang, J., and Yu, Y. Activation maximization generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, 2018.