Synthesis of Musical Instrument Sounds: Physics-Based Modeling or Machine Learning?

Scott H. Hawley

Address:
Department of Chemistry & Physics
Belmont University
1900 Belmont Boulevard
Nashville, Tennessee 37211
USA

Email:
scott.hawley@belmont.edu

Vasileios Chatziioannou

Address:
Department of Music Acoustics (IWK)
University of Music and
Performing Arts Vienna
Anton-von-Webern-Platz 1
1030 Vienna
Austria

Email:
chatziioannou@mdw.ac.at

Andrew Morrison

Address:
Department of Natural Sciences
Joliet Junior College
1215 Houbolt Road
Joliet, Illinois 60431
USA

Email:
amorriso@jjc.edu

Physics-based modeling provides insight into sound production processes, whereas machine learning generates increasingly realistic imitations from recordings alone.

Introduction

Music (and sound) synthesis has become widespread and is found across many musical genres. Many members of the Acoustical Society of America (ASA) also use various types of synthesis to produce the sounds used in research projects, ranging from measurements of hearing to studies of sound propagation in the oceans. Although there is a long history of attempts to synthesize music and other sounds (e.g., Pejrolo and Metcalfe, 2017), this paper focuses on the most recent, and very exciting, approaches now being used to synthesize the sounds of musical instruments as well as for a wide range of other acoustic phenomena. Earlier methods for musical instrument sound synthesis used signal-processing effects such as frequency-modulated (FM) synthesis or wavetables (i.e., sampling), whereas the applications of physics-based modeling and machine learning have only been applied relatively recently because of their higher computational cost.

Musical acoustics is a diverse scientific field that deals with research and applications ranging from making musical instruments to the perception of sound. One of the driving questions for musical acoustics is "How does this musical instrument produce its characteristic sound?" Well-known scientists throughout history, including the ancient Greek mathematician and philosopher Pythagoras, Galileo Galilei, Ernst Chladni, Herman von Helmholtz, and Chandrashekhara Venkata Raman (Nobel Laureate in Physics), increased the overall scientific understanding of their eras by working on music and musical instruments. The formulation of reliable physical models of musical instruments was pursued after developments in the field of differential equations, whereas digital sound synthesis was only possible following advances in numerical analysis and computer science.

The ability to digitally model musical instruments provides music creators with multiple desirable capabilities (Smith, 2011), among which are

(1) portability: virtual instruments and software effects require no space or weight;
(2) flexibility: many such instruments can be stored and accessed together and quickly modified;
(3) signal to noise: often can be higher with digital instruments;
(4) centralized, automated control;
(5) repeatability: simulated instruments can be exactly the same as opposed to physical systems that may involve variations due to, for example, construction, humidity, and temperature; and
(6) extension: the development of digital instruments involves fewer constraints than their real-world counterparts.

https://doi.org/10.1121/AT.2020.16.1.20

The process of constructing such virtual models has traditionally been performed using one of two main approaches. One approach is to construct a physics-based model (often through computer simulation) of the processes that take place during sound generation by a musical instrument. For example, a simulation may seek to predict the vibrations of the parts of an instrument, including air fluctuations, as well as all of the physical processes leading to the radiation of the sound from the instrument. Some physics-based modeling has taken the approach of finding electrical circuit analogs to the mechanical system under study.

The other main sound synthesis approach has been to emulate the sonic features of the instrument via signal-processing techniques. The synthesis of instrument sounds in this approach is done with the intent to capture the salient aspects of sounds produced by the instrument(s) under consideration, without regard for how they may have originated. A long-standing example of this is wavetable-based synthesis, in which the frequency features of an instrument are modulated in time by transients such as attack, decay, sustain, and release (ADSR).

In recent years, a significant new tool to accomplish the aforementioned synthesis of acoustic sounds is that of machine learning (ML). ML generally refers to the use of data analysis to discover patterns from large datasets such that a predictive model can be iteratively "trained" to produce outputs that increasingly approximate the desired results. Although many various forms of ML exist, some of the most powerful ML methods have employed artificial neural networks, which can be regarded as a set of curve-fitting approximation methods that use a series of "layers" of matrix multiplications with nonlinear functions applied between each matrix operation. When there are many layers, the model is called a "deep" neural network and its training is called "deep learning" (which is regarded as a subset of ML). Deep-learning methods have become increasingly used in a wide variety of research fields, including astrophysics, genetics, engineering, and acoustics, for tasks such image labeling, automated data acquisition, and speech recognition, often rivaling or at times besting the state-of-the-art methods previously crafted carefully by human experts. The use of deep learning for audio synthesis has been termed neural audio synthesis (NAS) and is discussed in **Neural Audio Synthesis**.

If one attends an ASA meeting, sessions in nearly every subdiscipline (e.g., bioacoustics, underwater acoustics) will feature talks and posters applying either physics-based modeling or deep learning to the specific acoustics domain featured in the session. Thus, although in this paper we focus on providing an update on these two approaches in the specific context of the synthesis of musical instrument sounds, such methods are generally applicable to other acoustics domains as well.

Physics-Based Modeling

The function of musical instruments has been studied by physicists, engineers, and scientists in general, not only due to the dominant role of the instruments in music production but also because of the compelling physical effects that govern the process of sound generation. An early attempt on physics-based modeling was presented by Kelly and Lochbaum (1962) focusing on speech synthesis by modeling the human vocal tract. Until the 1980s, various methods were developed that were suitable for musical instrument simulation, such as finite-difference models, mass-spring networks, and wave digital filters (reviewed in Välimäki et al., 2006).

Early digital sound synthesis methods were developed in the 1950s to generate audio/music in the absence of a musical instrument. These methods, such as FM synthesis, attempt to replicate sound spectra without being based on underlying physical laws. They are still popular in the field of electroacoustic composition, but they fail to offer realistic control over a digital musical instrument.

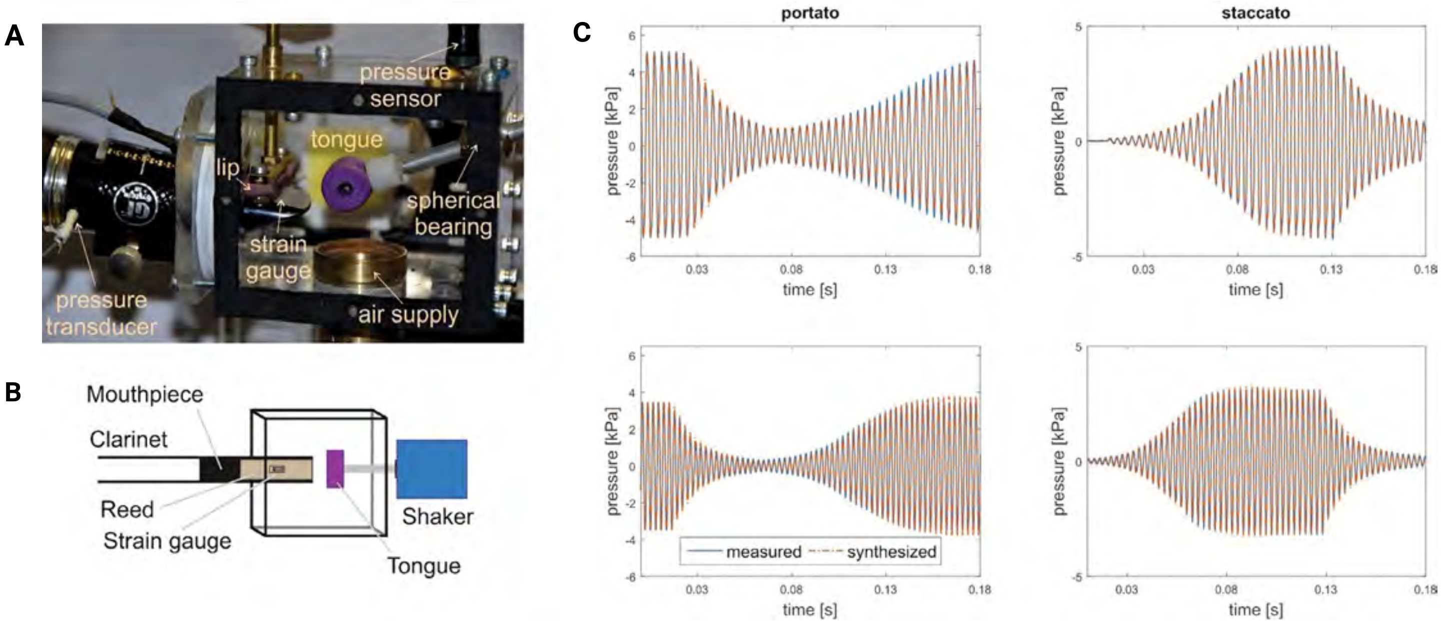
On the other hand, physics-based modeling aims to simulate the sound generation mechanism of musical instruments. Thus, physics-based modeling offers a physics-based reproduction of waveforms under both static and dynamic conditions, the possibility to model transient and nonlinear phenomena, and intuitive control over the involved physically meaningful parameters. Sounds generated by physics-based modeling can contain all the subtle audio information of sounds produced by a real instrument. Furthermore, physics-based modeling presents the possibility to estimate model parameters from naturally performed sounds. These parameters can be associated with certain playing techniques and may therefore be used to reveal differences not only across instruments but also across instrumentalists.

Detailed studies are now available on the function of musical instruments (e.g., Fletcher and Rossing, 1998). These studies also include analysis of complex nonlinear phenomena that attract a great deal of attention in musical acoustics, such as the excitation mechanism in woodwind instruments (Wolfe, 2018). Time-domain simulations of musical instruments are facilitated by solving the (partial) differential equations that describe their oscillations. Several time-stepping algorithms exist for the numerical treatment of such equations in the time domain. Most of the relevant algorithms found in the musical acoustics and sound-computing literature are based on finite differences or closely related methods (Välimäki et al., 2006). When nonlinear phenomena need to be modeled, stability can be shown for some specific cases or under specific assumptions. However, this is not always the case, such as when dealing with systems where the underlying physical quantities are nonanalytic functions of the phase space variables (Chatziioannou and van Walstijn, 2015). As such, numerical analysis techniques are still under investigation in the simulation community (Bilbao et al., 2015).

Regarding the actions of the player, numerical simulations are still rather limited because most physics-based models usually neglect the fact that the instrument oscillations can be excited in different manners during a virtuosic performance (Wolfe et al., 2015). To the contrary, idealized initial and boundary conditions are often employed in numerical simulations, which do not reflect the control of the player on the instrument. Studying the continuous interaction between the player and the instrument (inherently present in woodwind instruments) may drive forward the formulation of physical models, enabling the modeling of a wide variety of playing gestures. We will use the case of single-reed woodwind instruments (e.g., clarinets and saxophones) as an example of player-instrument interactions.

During expressive woodwind performance, musicians use various articulation techniques, mostly involving different kinds of interaction between the players' tongues and the vibrating reed (Scavone, 1996). To capture such interactions, an additional term has been included in the equations that model the single-reed excitation mechanism corresponding to tongue-reed interaction (Chatziioannou et al., 2019). This nonlinear term is added to two further nonlinearities that take place at the driving end of the instrument, namely, the collision between the reed and the mouthpiece (Dalmont et al., 2003) and the flow that enters the mouthpiece through the reed (Wolfe, 2018). Thus the force acting on the oscillating reed can be written as $f_{\text{tongue}} + f_{\text{lay}} + f_{\Delta}$, where f_{tongue} is the force of the player's tongue acting on the reed, f_{lay} is the collision force between reed and mouthpiece, and f_{Δ} is the force due to the pressure difference across the reed. This excitation model may be coupled to a linear model that describes wave propagation inside a cylindrical tube, which is an accurate approximation in the case of the relatively short bores of woodwinds. The coupled model may be solved numerically using, for example, the finite-difference

Figure 1. ***A:*** *artificial blowing machine for single-reed woodwind instruments.* ***B:*** *sketch of the blowing machine (view from above).* ***C:*** *examples of measured and synthesized mouthpiece pressure for staccato (an individual note separated from its neighboring notes by silence) and portato (the notes are generally sustained, using tonguing to achieve some separation) articulation.*



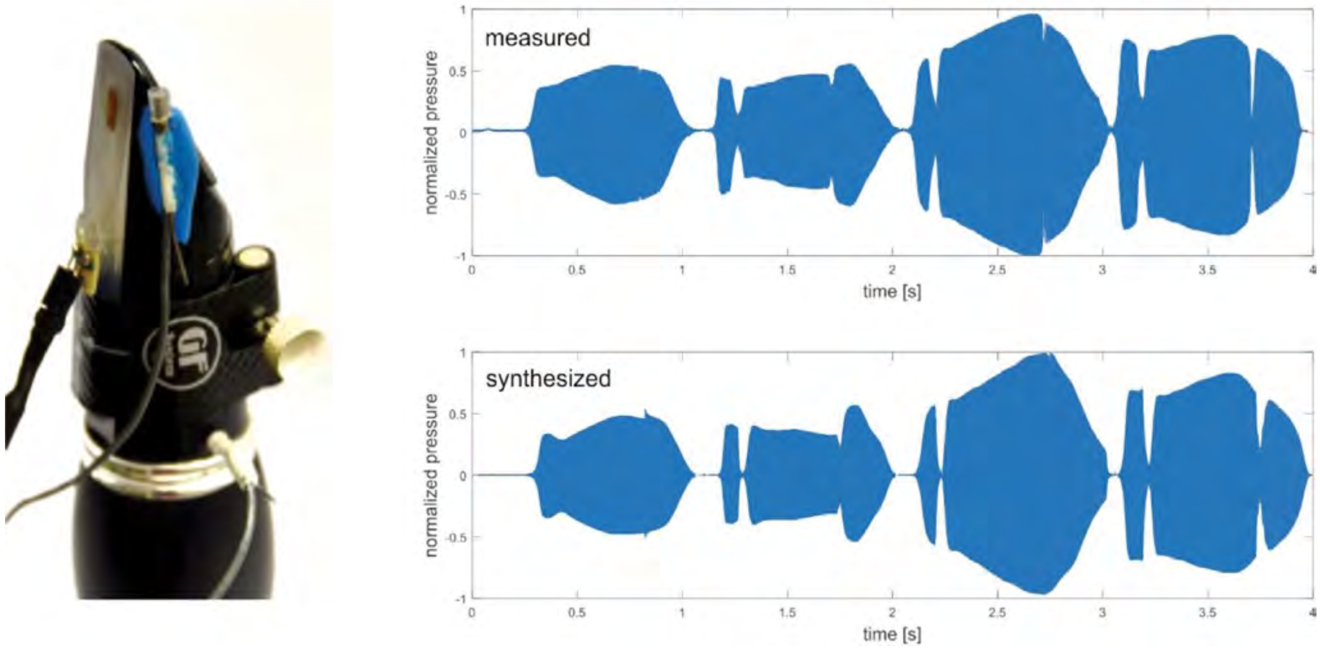


***Figure 2. Left:** instrumented clarinet mouthpiece for measurements with human players. Two pressure transducers are used to measure the pressure inside the instrument (mouthpiece pressure) and inside the player's mouth (blowing pressure). A strain gauge is used to monitor the vibrations of the reed. **Right:** measured and synthesized mouthpiece pressure for an excerpt from Weber's Clarinet Concerto no. 2 (see also **Multimedia1** and **Multimedia2** at acousticstoday.org/hawleymedia).*

method and yields information regarding the pressure and flow inside the tube, whereas the radiated pressure and the reed state are also calculated.

To validate this player-instrument interaction model, the numerically synthesized signals have been compared to measurements using (1) an artificial blowing machine and (2) experiments carried out with human players. The blowing machine shown in **Figure 1** establishes the conditions for repeatable laboratory measurements. The mouthpiece pressure obtained via the blowing machine is compared with the one calculated using a physical model for different articulation conditions (Chatziioannou et al., 2019). The success of this resynthesis process points toward the fact that all significant physical phenomena that take place during such note transitions are accurately captured by the physical model.

Figure 2 shows a similar comparison between human players and physics-based modeling. The blowing pressure and the mouthpiece pressure are measured during a performance using two pressure transducers (one inside the player's mouth and one inside the mouthpiece), whereas the reed displacement is monitored by means of a strain-gauge sensor. The model is capable of qualitatively reproducing an excerpt performed by a professional musician (taken from the Clarinet Concerto no. 2 by Carl Maria von Weber; see **Multimedia1** and **Multimedia2** at acousticstoday.org/hawleymedia).

During the above inverse-modeling applications, the fact that the involved model parameters have a physical nature allows intuitive access to them in terms of both control and interpretation (Campbell et al., 2004). Hence, it is possible to analyze the function of an instrument as well as the control exerted on it by the player by studying how the parameters of the physical model vary during a performance. Such physics-based analysis requires the formulation of accurate models and the design of suitable experimental setups to determine those parameters that influence the sound generation mechanism. Deep neural networks have also been used for parameter estimation and sound resynthesis (Gabrielli et al., 2018), but their application has been limited on isolated notes and yielded signals that are only qualitatively similar to the recorded ones. Such networks can be used for more than parameter estimation; in **Neural Audio Synthesis**, we discuss synthesizing waveforms directly using neural networks.

Neural Audio Synthesis

In contrast to physics-based modeling, NAS approaches seek to minimize the difference between a recorded audio sound (or set of sounds) from a musical instrument and audio synthesized by a deep neural network. This minimization process is carried out by means of some (typically gradient-based) optimization procedure applied to some metric or *loss* function. Thus, the development of NAS models obviates the need for explicit physics-based models, instead requiring a "training dataset" of prerecorded audio sounds against which to refine the model's outputs. Rather than relying on physically relevant control parameters, NAS models may "learn" alternative parameterizations of the sound.

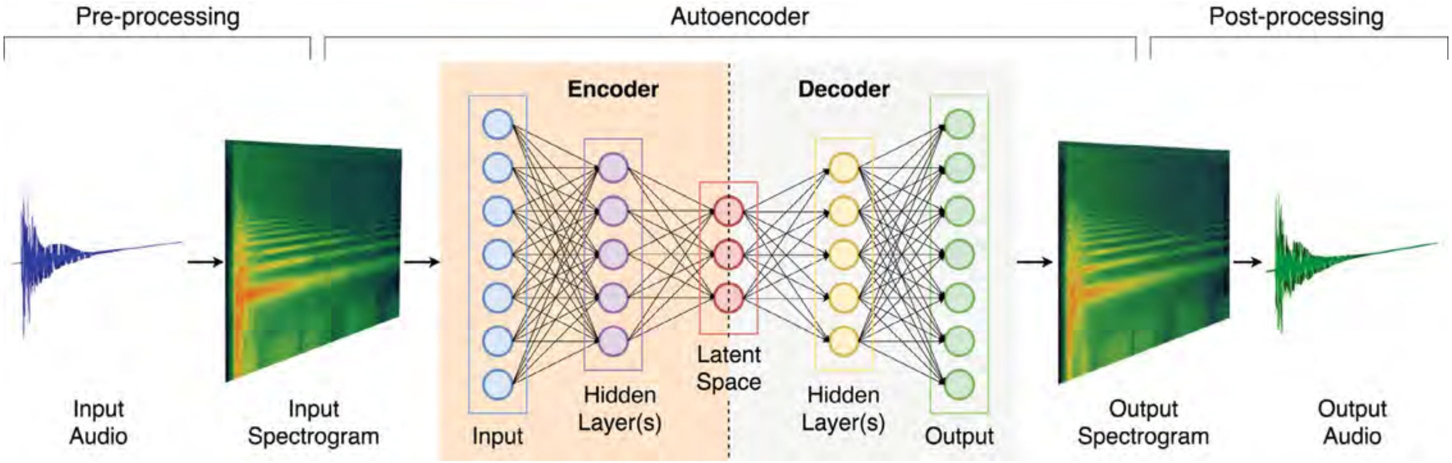
NAS architectures have primarily taken the form of autoencoders or Generative Adversarial Networks (GANs), which we describe shortly. An attractive feature of these two architectures is that they are "unsupervised" (or "self-supervised") learning models for which the algorithm does not require human labeling of their datasets, an enterprise that can suffer from expenditures of time and finances or concerns about accuracy.

Autoencoder Approaches

An autoencoder is a deep neural network trained to reproduce its input. It consists of multiple layers of artificial neurons arranged in an encoder-decoder pair (illustrated in **Figure 3**). The "hourglass" shape of the autoencoder forces the model to "learn" a compressed parameterization, often referred to as a "latent space representation" (i.e., **Figure 3**, *center* of hourglass), for determining the output signal. This reduced set of encoded features can then be altered slightly and decoded to synthesize new forms of audio, that is, one can later use the decoder portion alone as a synthesizer. The encoder and decoder may consist of one or multiple layers of neural connections, which can allow differing hierarchies of modeling complexity (see Roche et al., 2018, for a comparison). The inputs and outputs of the autoencoder may be raw audio waveforms, so-called "end-to-end" models, but more typically are magnitude spectrograms obtained via short-time Fourier transforms (STFTs) or related transformations. To produce the final waveform from the output spectrogram, there are well-known iterative techniques that can be used (Griffin and Lim, 1984). The hidden layers within the encoder and decoder may be simple "fully connected" layers (i.e., matrix multiplications followed by nonlinear activation potentials) or may involve structures that allow for more efficient capturing of behavior over "long" timescales, such as recurrent neural network layers with an internal memory (Mehri et al., 2016), or a stacked series of dilated convolutions as in the WaveNet scheme (van den Oord et al., 2016).

A noteworthy end-to-end autoencoder model known as NSynth was created in the Google Magenta group (available at tinyurl.com/gmagenta; Engel et al., 2017). This group trained two different autoencoder models on an extremely large dataset of musical instrument sounds consisting of "~300k four-second annotated notes sampled at 16 kHz from ~1k harmonic musical instruments." They compared the performance of a baseline spectral autoencoder model of the type described above with an autoencoder that used a WaveNet structure, finding the latter offering significant improvements in reproducing aspects of tone quality, attack transients, timbre, and dynamics. The latent-space encod-

Figure 3. *Schematic of an autoencoder method in which the output spectrogram of a neural network approximates its input spectrogram. Here we show fully connected neural network layers operating on spectrograms, whereas other autoencoders make use of more complex network architectures (e.g., recurrent or convolutional neural network layers) and operate directly on raw audio.*



ing of these attributes provided an opportunity to merge multiple instrument sounds, such as "bass + flute" or "flute + organ." The ability for NSynth to merge sounds allowed for the creation of a physical touchpad controller (Engel, 2017) from which musicians could interpolate between sounds and generate new combinations in real time. Beyond the utility of the NSynth model itself, the NSynth dataset has provided significant benefit for the field because other research groups have used this dataset for training other models both to try to exceed the performance of the NSynth model (e.g., Défossez et al., 2018) and as a means of baseline comparison between different models (e.g., Roche et al., 2018).

As noted above, the latent space representation in an autoencoder can be altered and decoded to synthesize new sounds; however, different types of sounds may become associated with disjoint regions of the latent space, making interpolation between instruments behave strangely and produce unexpected results. Furthermore, the decoding of a given set of latent features is identical each time. To provide variety in the instrument sounds and recast the system as a truly "generative model" (i.e., one that produces novel output on each use), the autoencoder paradigm can be altered to model the *probability distribution* of output audio features as a function of a learned probability distribution of features in latent space. Such systems are known as a variational autoencoders (VAEs). Although on a simplified level VAEs often amount to replacing single values in the latent space with the mean and standard deviation of a Gaussian distribution, VAEs can be considerably more difficult to train than ordinary "vanilla" autoencoders, and significant VAE results for musical instrument synthesis have only appeared relatively recently (e.g., Çakir and Virtanen, 2018). VAEs form a bridge to another class of generative models that also model probability distributions of sounds but dispense with the autoencoder form, as follows.

Generative Adversarial Network Approaches

A powerful paradigm emerging in recent years for the generation of synthetic data is the GAN. A GAN can be regarded as two competing deep neural networks whose efforts are combined in a kind of "arms race" (illustrated in **Figure 4**). One part of the network is called the "generator" that synthesizes new data; the other part of the network is the "discriminator" that functions as a classifier to determine whether the data at its input is coming from the generator or is prerecorded data. This process has been likened to the process of counterfeiting, where the generator is the "criminal artist" and the discriminator is the "forensic detective." The output from the discriminator is used to train the generator, so that over time, its outputs more closely resemble the prerecorded audio.

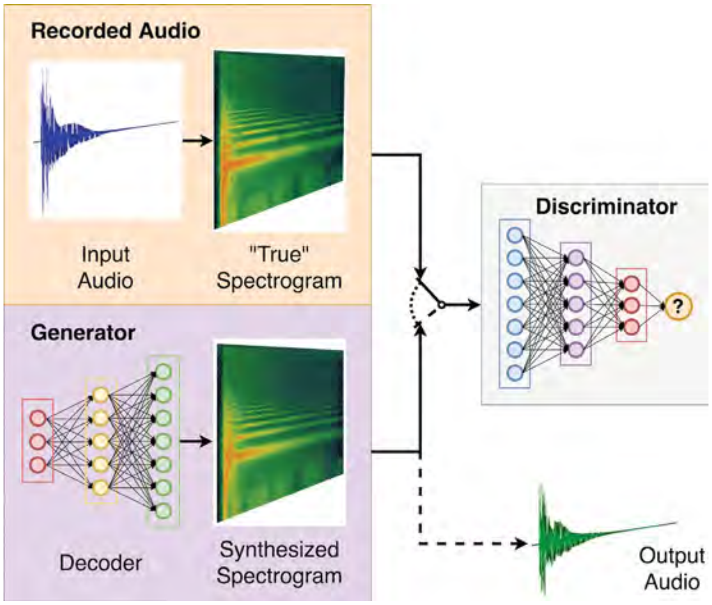


Figure 4. *Overview of a generative adversarial network (GAN), a sort of "imitation game" played between two neural networks: a binary classifier called the discriminator seeks to improve at correctly "guessing" whether its input came from the dataset of instrument recordings or is a "forgery" synthesized by the generator. The generator uses information from the optimization procedure of the discriminator (e.g., the negative of the gradients) to synthesize increasingly "convincing" instrument sounds.*

Initially, GANs were applied to image synthesis, followed by some synthesis of speech audio, but musical instrument synthesis remained untouched until a noteworthy preprint appeared in early 2018, stating "In this paper we introduce WaveGAN, a first attempt at applying GANs to unsupervised synthesis of raw-waveform audio" (Donohue et al., 2019). WaveGAN applied a one-dimensional version of the two-dimensional convolutions used for image-synthesizing GANs (i.e., it did not use a WaveNet architecture) applied to raw audio samples. The paper featured another model, "SpecGAN," operating on spectrogram images in the manner of preexisting GANs. These two models were trained on datasets that included piano and drums. In the case of drums, the WaveGAN model was featured in a Web demo of an interactive drum machine (see **Multimedia3** at acousticstoday.org/hawleymedia) for which novel drum samples could be synthesized at the press of a button. Although SpecGAN tended to produce slightly more easily

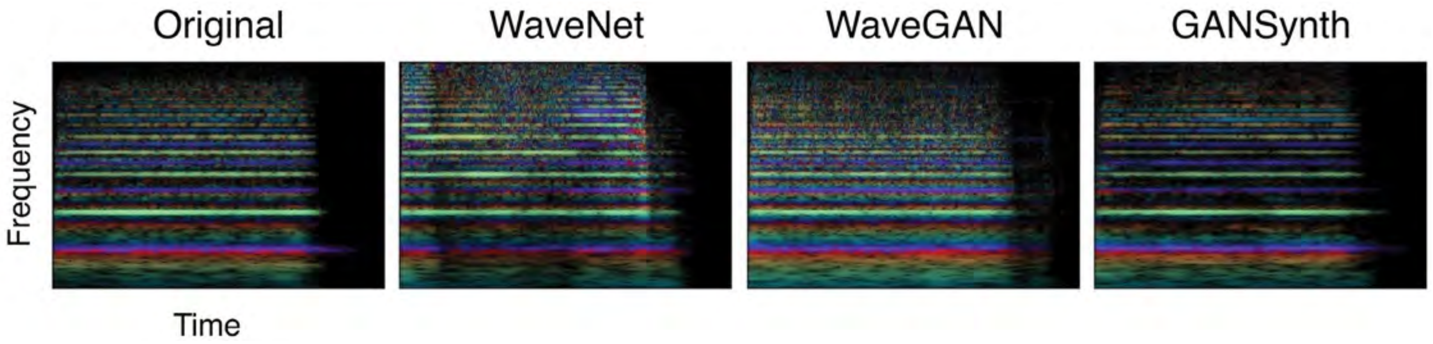


Figure 5. *Synthesis results displayed as "rainbowgrams" (after Engel et al., 2017), showing the intensity given by the logarithm of the magnitude, with frequency scaled logarithmically and colored by the instantaneous frequency. Shown are similar tones for a sustained C3 note of a string bass. The original recording and GANSynth example show greater color consistency, corresponding to greater phase coherence and perceived tone quality, than the WaveNet and WaveGAN examples.*

identifiable sounds for human speech, WaveGAN was preferred by human listeners for its sound quality and diversity.

Combining the successes of NSynth and WaveGAN along with new refinements and insights, two new similar models known as GANSynth (Engel et al., 2019) and TiFGAN (Marafioti et al, 2019) now stand as the state of the art for the NAS of musical instrument sounds. One improvement resulted from increasing the number of frequency bins in the STFT outputs via overlapping the frames more. Another insight used in these methods is the use of "instantaneous frequency," which is the time derivative of the phase produced by the STFT, in the neural network. **Figure 5** shows a comparison between outputs from the WaveNet autoencoder used in the NSynth release, WaveGAN, and GANSynth. Given the (random) generative nature of these models, an exact tonal comparison is not feasible; however, we show similar-sounding tones for the C3 note (130.81 Hz) corresponding to the sound of a bass string. (For additional audio examples, click the Audio Examples link on the GANSynth Web page available at tinyurl.com/gansynth). The GANSynth output is superior to previous synthesized outputs in terms of its phase coherence, indicated in these "rainbowgrams" for which instantaneous frequency determines the color. Beyond quantitative testing, in a series of human trials consisting of 3,600 listening evaluations, listeners preferred GANSynth results, with approval ratings approaching (i.e., within 10% of) their preferences for acoustic recordings.

Unlike its WaveNet-based predecessors, which were developed for variable-length audio sources such as speech and thus tended to be autoregressive (i.e., predicted one sample at a time using previous outputs) and therefore slow, GANSynth generates entire audio clips all at once. Thus, both training the model and generating new samples occur much faster than previous methods, making it attractive for real-time sample generation. However, it remains to be seen how well it can be adapted for variable-length outputs such as entire music compositions (instead of single notes).

Although tone quality has seen significant improvement and nears acoustic recordings in listening tests, the audio is typically generated using lower sample rates than compact disc (CD) quality sound. For example, the NSynth dataset used a sample rate of 16 kHz, which allows for a maximum representable frequency of only 8 kHz. The use of such reduced sample rates is a common theme in the NAS community, with CD-quality sample rates found in greater occurrence among specific domains such as audio engineering (e.g., Hawley et al., 2019). For many musical instruments, however, the magnitude of spectral content above 8 kHz is typically several orders of magnitude below lower frequency sounds so that the use of reduced sample rates remains a reasonable restriction for these use cases. Physics-based modeling, on the other hand, can generate tones at arbitrary sample rates and often without the issues of noise and phase coherence associated with the iterative approximation scheme of NAS methods.

Conclusions

This article provides an update on the methods for musical instrument sound synthesis. Depending on the intended goal of the synthesis and the resources at one's disposal, a variety of methods are available, among which are detailed physics-based modeling of a musical instrument as well as NAS (i.e., the use of deep neural networks for producing sounds that approximate the salient features of the tonal qualities of an instrument). Physics-based modeling relies on solving equations comprising a mathematical model of the processes at work in an instrument and the body of the musician playing it. Beyond the application of tone production alone, physics-based modeling can be used to provide specific and detailed predictions that can be compared with real phenomena to test the validity and relevance of the physical model itself or to determine whether additional phenomena need to be included in the model. This can contribute to a greater understanding in the field of musical acoustics overall.

A large class of instances of "applied" musical acoustics use cases involves only the production of instrument sounds, for example, as virtual instruments for musicians and composers. In these use cases, "convincing" sounds need not be derived from detailed physics simulations. Methods for synthesizing musical instrument sounds from a reduced set of salient factors have been available for decades and are continually improving in quality; however, these are usually carefully crafted by human experts with domain-specific knowledge. An alternative approach seeing increasing popularity and success in recent years is to have a deep neural network "automatically learn" its own representation by using only a large corpus of audio recordings. Creating a NAS-based instrument model may not yield physical insight but also requires none and thus can be developed by those lacking domain-specific physical insight. Such models may require hundreds of hours of recorded audio and days of computation to train. However, once trained, they can execute quickly enough to be used in real time to generate arbitrary numbers of variations without requiring a new physics simulation of all processes involved.

Both physics-based modeling and NAS provide viable pathways to synthesizing realistic musical instrument sounds, and the choice of which approach to employ will likely depend on the experience and preferences of the developers. Currently, we are unaware of a direct comparison of sound quality between neural-generated and physically modeled sounds; however, this is an avenue we hope to see more of in the future. Also, we are intrigued by the possibility of developers using physically modeled sounds to train neural network systems. Such an approach would mirror other areas of physics in which neural network systems trained on large-scale physics simulations (e.g., of earthquake propagation) are being used to produce suitable approximate results for novel system parameters at a fraction of the time and expense needed to rerun a full physical model. Thus, it is evident that physics-based modeling and NAS approaches can be seen as complementary methods to advance the field of musical acoustics.

References

Bilbao, S., Torin, A., and Chatziioannou, V. (2015). Numerical modeling of collisions in musical instruments. *Acta Acustica united with Acustica* 101, 155-173. Available at https://tinyurl.com/btznum.

Çakir, E., and Virtanen, T. (2018). Musical instrument synthesis and morphing in multidimensional latent space using variational, convolutional recurrent autoencoders. *Proceedings of the Audio Engineering Society 145th International Conference*, New York, October 17-19, 2018. Available at https://tinyurl.com/morphae.

Campbell, M., Greated, C., and Myers, A. (2004). Musical Instruments: *History, Technology, and Performance of Instruments of Western Music.* Oxford University Press, Oxford, UK.

Chatziioannou, V., Schmutzhard, S., Pàmies-Vilà, M., and Hofmann, A. (2019). Investigating clarinet articulation using a physical model and an artificial blowing machine. *Acta Acustica united with Acustica* 105, 682-694. Available at https://tinyurl.com/chatz19.

Chatziioannou, V., and van Walstijn, M. (2015). Energy conserving schemes for the simulation of musical instrument contact dynamics. *Journal of Sound and Vibration* 339, 262-279. Available at https://tinyurl.com/chatz2015.

Dalmont, J.-P., Gilbert, J., and Ollivier, S. (2003). Nonlinear characteristics of single-reed instruments: Quasistatic volume flow and reed opening measurements. *The Journal of the Acoustical Society of America* 114(4), 2253-2262. Available at https://tinyurl.com/dalmont03.

Défossez, A., Zeghidour, N., Usunier, N., Bottou, L., and Bach, F. (2018). SING: Symbol-to-Instrument Neural Generator. *Conference on Neural Information Processing Systems* (*NeurIPS*), Montreal, QC, Canada, December 2-8, 2018. Available at https://tinyurl.com/defossez.

Donahue, C., McAuley, J., and Puckette, M. (2019). Adversarial audio synthesis. *International Conference on Learning Representations*, New Orleans, LA, May 6-9, 2019. Available at https://tinyurl.com/waveganpaper.

Engel, J. (2017). *Making a Neural Synthesizer Instrument*. Available at https://tinyurl.com/nsynthbuild.

Engel, J. H., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019). GANSynth: Adversarial neural audio synthesis. *Computing Research Repository*, arXiv:1902.08710.

Engel, J. H., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. (2017). Neural audio synthesis of musical notes with WaveNet autoencoders. *Proceedings of the 34th International Conference on Machine Learning-Volume* 70, Sydney, NSW, Australia, August 6-11, 2017, pp. 1068-1077.

Fletcher, N., and Rossing, T. (1998). *The Physics of Musical Instruments*. Springer-Verlag, New York.

Gabrielli, L., Tomassetti, S., Zinato, C., and Piazza, F. (2018). End-to-end learning for physics-based acoustic modeling. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2), 160-170. Available at https://tinyurl.com/gabrielli18.

Griffin, D., and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 236-243. Available at https://tinyurl.com/griffinlim.

Hawley, S. H., Colburn, B. L., and Mimilakis, S. I. (2019). Profiling audio compressors with deep neural networks. *Proceedings of the Audio Engineering Society 147th International Conference,* New York, October 16-19, 2019. Available at https://tinyurl.com/signaltrainpap.

Kelly, J., and Lochbaum, C. (1962). Speech synthesis. *Proceedings of the 4th International Congress on Acoustics*, Copenhagen, Denmark, August 21-28, 1962.

Marafioti, A., Perraudin, N., Holighaus, N., and Majdak, P. (2019). Adversarial generation of time-frequency features with application in audio synthesis. *Proceedings of the 36th International Conference on Machine Learning* 97, Long Beach, CA, June 10-15, 2019, pp. 4352-4362. Available at https://tinyurl.com/marafioti.

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., and Bengio, Y. (2016). SampleRNN: An unconditional end-to-end neural audio generation model. *Computing Research Repository*, arXiv:1612.07837.

Pejrolo, A., and Metcalfe, S. B. (2017). *Creating Sounds from Scratch: A Practical Guide to Music Synthesis for Producers and Composers*. Oxford University Press, Oxford, UK.

Roche, F., Hueber, T., Limier, S., and Girin, L. (2018). Autoencoders for music sound synthesis: a comparison of linear, shallow, deep and variational models. *Sound and Music Computing Conference*, Malaga, Spain, May 28-31, 2019. Available at https://tinyurl.com/rochepdf.

Scavone, G. P. (1996). Modeling and Control of Performance Expression in Digital Waveguide Models of Woodwind Instruments. *Proceedings of the 1996 International Computer Music Conference*, Hong Kong, August 19-24, 1996. Available at https://tinyurl.com/scavone96.

Smith, J. O. (2011). *Physical Audio Signal Processing: For Virtual Musical Instruments and Audio Effects.* Free ebook by Julius Smith, Stanford Center for Computer Research in Music and Acoustics, W3K Publishing. Available at https://ccrma.stanford.edu/~jos/pasp/.

Välimäki, V., Pakarinen, J., Erkut, C., and Karjalainen, M. (2006). Discrete-time modelling of musical instruments. *Reports on Progress in Physics* 69, 1-78.

van den Oord, A., Dieleman, Zen, S. H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *Computing Research Repository*, arXiv:1609.03499.

Wolfe, J. (2018). The acoustics of woodwind musical instruments. *Acoustics Today* 14(1), 50-56.

Wolfe, J., Fletcher, N., and Smith, J. (2015). The interactions between wind instruments and their players. *Acta Acustica united with Acustica* 101, 211-223.

BioSketches

Scott Hawley received his PhD in numerical relativity from the University of Texas at Austin. After two postdoctoral positions simulating black holes, his interest in music led to a teaching position at Belmont University in Nashville (TN). As a professor of physics teaching acoustics to audio engineering majors, he has developed sound visualization apps for education and machine-learning software for music industry professionals. These include the iOS app Polar Pattern Plotter that was featured on the cover of *The Physics Teacher*. His Vibrary neural network-based music information retrieval system for composers was the winner of the 2018 Incubator Lab Development Award from Art+Logic Inc.

Vasileios Chatziioannou received his PhD in the field of electronics, electrical engineering, and computer science from Queen's University Belfast. Subsequently, he moved to the Department of Music Acoustics at the University of Music and Performing Arts Vienna where he is teaching and conducting research in physical modeling of musical instruments and virtual reality audio. He has led the *Transient Phenomena in Single-Reed Woodwind Instruments* project, funded by the Austrian Science Fund (FWF), and the ITN project *VRACE — Virtual Reality Audio for Cyber Environments*, funded by the Horizon 2020 Framework Programme of the European Union.

Andrew Morrison received his PhD in the field of musical acoustics from Northern Illinois University (DeKalb). He has taught physics since 2011 at Joliet Junior College, the country's oldest two-year college, located in Joliet, IL, just outside Chicago. He served as chair of the Technical Committee on Musical Acoustics for the Acoustical Society of America from 2014 to 2019 and currently serves as a member of the Acoustical Society of America Executive Council.