forum acusticum 2023

TOWARDS ORCHESTRATING PHYSICALLY MODELLED 2D PERCUSSION INSTRUMENTS

Lewis Wolstanholme
Centre for Digital Music
Queen Mary University of London
United Kingdom

Andrew McPherson
Dyson School of Design Engineering
Imperial College London
United Kingdom

ABSTRACT

Contemporary orchestration practice harbours a number of aesthetic inquiries relating to the employment and arrangement of percussion instruments. Due in part to the fact that percussion instruments largely occupy an inharmonic timbre space, they encompass a diverse and distinctly nuanced musical idiom in comparison to harmonic instruments, both in terms of their textural interplay, musical function and cultural significance. In response to this perspective, we present a neural network approach to the parameter estimation of physically modelled, abstract percussion instruments. The approach presented here serves as our initial attempt towards creating a computer-assisted orchestration methodology specifically targeting the musical employment and arrangement of inharmonic timbres and percussive instruments. The neural architecture presented here has been trained and tested using a pair of two-dimensional physical models, to gauge a sense of our architecture's successes and limitations as we continue to expand this approach to include more two-dimensional models. This works poses as our first technological inquiry into this field, which has here been quantitatively assessed, with plans to undertake more rigorous and comparative tests in the near future.

Keywords: *parameter estimation, orchestration, percussion, artificial intelligence*

1. INTRODUCTION

Percussion instruments play a crucial role within many cultural idioms, and yet their timbral prominence in musical discourse is often difficult to discuss analytically and prescriptively. Orchestration, which is "one of the hardest musical disciplines to define and transmit" [1, p.99], is based upon many traditional approaches which struggle to find words beyond the term 'noise' and other textural descriptions to describe the timbral effects of percussion instruments [2, 3]. Although more contemporary approaches to orchestration highlight this 'noise-maker' notion as antiquated [4, p.474], much of the discourse is still limited in its scope in so far as timbral and spectral directives are concerned. In general, the detailed language and understanding that we have surrounding orchestration and arrangement with harmonic instruments does not enable detailed aesthetic explorations using the vast corpora of inharmonic and percussive sounds composers and arrangers have at their disposal.

This paper presents a neural network approach towards the orchestration of percussion instruments centred around estimating the parameters of physically modelled, two-dimensional percussion instruments. So far, we have developed a range of datasets focusing on specific parametric subsets of our synthetic percussion models, each of which explores both the instruments' geometry and size. We have also trained a convolutional neural network (CNN) on circular and rectangular membrane instruments, designed to infer their size and aspect ratio. These tools serve to both supply abstract sonic references to a wide variety of percussive instruments and timbres, and allow for the prediction of these sounds given an arbitrary target or reference sound. This approach to computational orchestration serves as a strong baseline from which we intend to expand our designs to incorporate more arbitrarily shaped drums and physical models in the future.

2. CULTURAL BACKGROUND

Despite the breadth of timbral diversity amongst percussion instruments, there has yet to emerge a methodology

through which percussion instruments can be employed specifically to create determinate timbral palettes - to form specific spectral colours by intentionally coalescing inharmonic instruments amongst themselves or alongside other harmonic instruments. In a practical sense, it is not yet easy to predict the degree to which a particular percussion instrument will compliment the sounds around it. The tuning of a snare, for example, is often decided from an isolated reference point, focusing on the sonic qualities of the snare itself, rather than from a place of understanding how a particular tuning may compliment other concurrent spectral and harmonic colours. The closest that western cultural practices come to this timbral understanding is through the idea of 'tuned percussion', which is largely influenced by an anterior musical discourse surrounding harmonic instruments, and often centres around percussive instruments having their timbral content skewed to more closely replicate the harmonic series [5, 6]. This notion is not strictly the case across all cultural practices, with many examples of Indonesian gamelan ensembles tuning their percussion instruments to instead complement one another [7]. In these cultural contexts, inharmonic spectra form the basis of their timbral and musical languages, often resulting in ensembles of percussion instruments that gradually reform and refine their timbral palettes over many generations [8].

Many of the theories underpinning this work originate from, and are reinforced by, our past discovery led [9] and practice based research projects, including our previous work *terracotta* [10] amongst others [11]. Whilst composing the work *terracotta*, a great deal of care was given to orchestrating the sounds of physically modelled, abstract percussion instruments. Driven solely by aural technique, each synthetic percussion instrument was tuned in relation to the other percussive sounds surrounding it, borrowing from the technique previously highlighted in Indonesian gamelan practices. After completing this work, we sought to challenge this approach to tuning and arranging abstract percussion instruments through our still ongoing composition study [12]. One of the key findings that we have uncovered during this study, is that percussion instruments are intuitively composed with via cultural references, rather than relying on the intricate spectral qualities of the instruments themselves. In this sense, a percussion instrument is typically incorporated into a work if it satisfies a particular audiated reference. And similarly, instruments which do not have strong sonic references are typically seen as less aesthetically applicable. This notion is both a blessing and a curse, for it encourages us to create strong idiomatic and semantic relationships with our instrumental sounds, but it also slows our desires to innovate using unfamiliar and non-standardised sounds. If the goal is to explore the extensive sonic possibilities of percussion and inharmonic sounds, which it is here, then we cannot rely on reference alone to satisfy our aesthetic explorations.

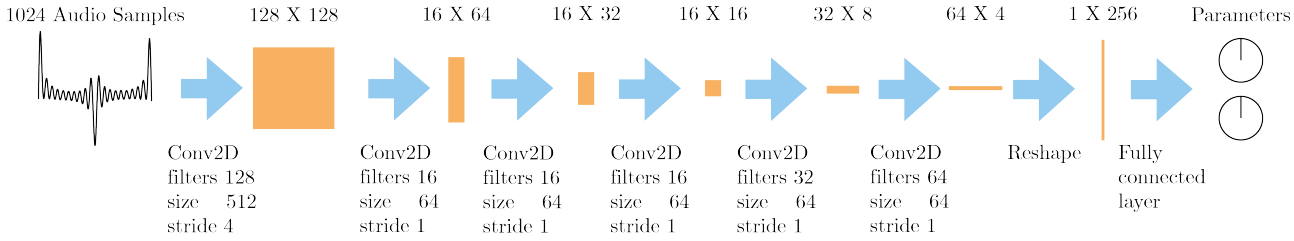
3. TECHNOLOGICAL BACKGROUND

For this work, our primary task is to create a neural network architecture that can perform parameter estimation on physically modelled, abstract percussion instruments. In terms of a more well defined 'behavioural objective' [13], we aim to create an end to end network that can predict continuous material properties of a physical model, based on an input audio sample. This would allow us to predict parameter information in response to arbitrary target sounds, such as pre-composed musical moments or reference sounds that we may aim to emulate, as well as perform parameter estimation based on constant audio streams. We are here focusing our efforts on a small subset of a physical model's parameters, namely its geometry and size. By enabling the prediction of a percussion instrument's material parameters in response to an arbitrary audio fragment, it will subsequently become possible to determine a complementary orchestral arrangement involving both the sounds contained within the audio fragment and the predicted instrument.

Parameter estimation networks have been developed to satisfy a range of use cases, from interfacing with VST synthesisers [14–16] and audio effects [17–19] to more large scale differentiable DSP tasks [20]. In some of the simpler use cases, parameter estimation has been achieved using a linear regression model for each individual parameter, which served to produce continuous estimations for each parameter in question [17, 18]. CNNs, on the other hand, have been shown to outperform linear regression models in a number of use cases [15, 19], however in both of the cases cited here, the objective had been translated from a regression task into a classification task. Although improved accuracy is a desired result, this redefinition of the task at hand does not support our architectural and behavioural aims. In some cases, CNNs have been implemented specifically to perform continuous parameter estimation of physically modelled instruments [21, 22], however the successes of these approaches were not evaluated in comparison to any other existing methodologies.

A similar and closely related field is the study of

Figure 1. The architectural design for our neural parameter estimation network.



target-based orchestration. Target-based orchestration is concerned with the transformation of a target sound, an arbitrary fragment of audio, into a full orchestral reconstruction represented using symbolic notation. This idea stems from the musical works of spectral composers, such as Gérard Grisey and Tristan Murail, with a software driven approach to orchestration originally being proposed back in the early 2000s by the composer and orchestration specialist Yan Maresz [1]. Throughout the development of numerous target-based orchestration systems, the predominant architectural approach has been to use a genetic algorithm [23–25], a methodology which is also used in the recent software orchestration tool *Orchidea* [26, 27]. A genetic algorithm is here used as a heuristic approach to solving a knapsack-like problem [28], asserting that the problem space be discretised and interpreted as a classification based problem. Although powerful, this methodology towards target-based orchestration is not directly applicable to our behavioural objective, due to its use of discretisation and classification. In recent years, however, there have been several attempts to reproduce this functionality using CNNs, as well as residual neural networks and long short-term memory networks [29, 30]. These models have not yet been shown to outperform *Orchidea*, however they do point towards promising results with respect to timbral representation and spectral orchestration.

4. NEURAL ARCHITECTURE

We base our neural network design on CRePE, a Convolutional Representation for Pitch Estimation [31]. As its full title suggests, CRePE was originally designed to work as a frequency estimation algorithm, to compete with other heuristic approaches to frequency estimation such as the PYIN [32, 33] and SWIPE [34] algorithms. Within the context of our use case, we took advantage of the fact that predicting the size of a percussion instrument is directly correlated with predicting its fundamental frequency. Additionally, CRePE has already been shown to perform accurate fundamental frequency estimation on inharmonic sounds [31, p.164], suggesting its architectural design is well suited for targeting the sounds of percussion instruments. Although these inclinations do not necessarily point towards a generalisable solution for the prediction of any physical model parameter, we believe that our CRePE inspired architecture will serve as a strong baseline for future work.

Similar to CRePE, we have developed a deep neural network with 6 convolutional layers and 1 fully connected layer,[1] as shown in figure 1. Each convolutional layer consists of a 2D convolution, followed by a ReLU activation function, batch normalisation and max pooling with a kernel size of 2. The input to our network is a tensor of 1024 audio samples, allowing our model to perform parameter estimation directly with raw audio waveforms without the need for preprocessing or the use of alternative time-frequency representations.

The biggest difference between our design and CRePE is our output layer. CRePE was originally designed to work as a classifier, and featured 360 output nodes, with each one representing a 20-cent interval between 32.7Hz and 1975.5Hz [31, p.162]. As a result, the loss function for this original architecture was calculated using binary cross entropy loss. In our work, we implemented a final layer which utilised a single output node per parameter. We calculated our loss in accordance with the mean squared error between our ground truth and predicted values, which was optimised using the Adam optimiser [35]. This approach allowed for the estimation of continuous parameter values, which could also, in theory, still be trained to estimate fundamental frequencies.

[1] This project's source code is available online: `https://github.com/lewiswolf/kac_prediction`

So far, we have only trained and deployed this model using one or two parameters at a time, and have not studied how well this architecture scales when incorporating large amounts of continuous output nodes. We instead aim to follow up on this specific inquiry in future works, as we continue to extend this model to incorporate more complex geometries with similarly complex numerical representations.

By its design, CRePE is an extremely large model, harbouring 22,239,976 trainable parameters. As part of the CRePE codebase[2], there is an option to downsize this number of parameters at the expense of decreased accuracy when deployed. After configuring our model in a similar way, and performing tests that assess the detriment of this reduced number of parameters, we did not observe a significant difference between there being a large or small number of them. For this reason, we chose to only work with a small number of parameters, the individual amounts for which are shown in figure 1. For a single output node, our model used a total of only 394,289 parameters, with an additional 257 parameters for each additional output node. Limiting ourselves to a small amount of parameters also enabled us to achieve improved performance during both training and deployment.

5. DATASET MODELLING

Our model was trained to optimise for two distinct tasks - finding the size of circular percussion instruments, and finding the size and aspect ratio of rectangular percussion instruments. To achieve these tasks, we trained our neural network on a dataset of 5000 circular drum sounds and a dataset of 5000 rectangular drum sounds, with each dataset being generated using a similar two-dimensional physical model [36]. Each dataset contained 1000 distinct drum sizes and geometries, with each one being sampled first in the centroid, and then in four random locations across the domain. These datasets were both split into training datasets, which were 70% of the original size, and testing and evaluation datasets, with each being 15% of the original size respectively.

Our physical model was based on modal synthesis techniques, cumulatively synthesising individual modal frequencies calculated in accordance with the two-dimensional wave equation. In this use case, the linear solution to the two-dimensional wave equation is calculated according to:

$$\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\cos(\omega_{mn}t)+\sin(\omega_{mn}t)\right)\alpha_{mn} \tag{1}$$

where m and n are the modal indices, and M and N are the maximum modal indices, here both chosen to be 10 [37, p.72].

For a rectangular domain with aspect ratio ϵ, the modal frequencies ω_{mn} can be calculated according to

$$\omega_{mn}=\pi\gamma\sqrt{\frac{m^2}{\epsilon}+\epsilon n^2} \tag{2}$$

where γ is the relative wavespeed (1/s). We calculate γ relative to the size of the drum in meters, L, the material density in kg/m^2, ρ, and the tension of the material in N/m, T, such that:

$$\gamma=\frac{\sqrt{T/\rho}}{L} \tag{3}$$

For the two datasets used here, the material density and tension were kept constant, at 0.2 kg/m^2 and 2000 N/m respectively, whilst the size of the models was randomly selected from within the range $0.1 \leq L \leq 2 \in \mathbb{R}$ and the aspect ratio was randomly selected from within the range $1 \leq \epsilon \leq 4 \in \mathbb{R}$. Under a Dirichlet boundary condition, the modal amplitudes α_{mn} can be calculated, relative to a cartesian strike location [38, p.309], according to:

$$\alpha_{mn}(x,y)=\sin\left(\frac{m\pi x}{\sqrt{\epsilon}}\right)\sin\left(n\pi y\sqrt{\epsilon}\right) \tag{4}$$

For a circular drum, it is more straightforward to notate the two-dimensional wave equation according to:

$$\sum_{m=1}^{M}\sum_{n=0}^{N-1}\left(\cos(\pi\gamma z_{mn}t)+\sin(\pi\gamma z_{mn}t)\right)\alpha_{mn} \tag{5}$$

Here, the modal frequencies ω have been replaced with $\pi\gamma z_{mn}$, where z_{mn} is used to represent the m^{th} positive zero crossing of an n^{th} order Bessel function of the first kind, such that $J_n(z_{mn})=0$ [37, p.74]. Again using a Dirichlet boundary condition, the modal amplitudes for a circular domain can be calculated, relative to a strike location in polar coordinates [39, p.210], according to:

$$\alpha_{mn}(r,\theta)=J_n(z_{mn}r)\left(\cos(n\theta)+\sin(n\theta)\right) \tag{6}$$

[2] The original source code for CRePE can also be found on GitHub: `https://github.com/marl/crepe`

6. EVALUATION

We initially tested our network using a random sweep of of the hyperparameter space. Each test was conducted with early stopping, which interrupted the training loop after 32 epochs without improvement. We tested both full batch and mini batch gradient descent, alternative optimisers such a stochastic gradient descent, as well as a range of values for both the dropout and learning rate. In conclusion, we found that the most optimal parameters for each task were the ones shown in table 1.

Table 1. Hyperparameters used when training our neural network on circular and rectangular percussion models.

	Circular	Rectangular
Batch Size	64	32
Dropout	0.25	0.4
Learning Rate	0.00025	0.005
Number of Epochs	267	171

Table 2. Mean squared error for our model when predicting the size of circular and rectangular percussion models, as well as the aspect ratio of rectangular models.

	Circular	Rectangular
Aspect Ratio	N/A	0.057
Size	0.001	0.047
Aggregate	0.001	0.033

After optimising the hyperparameters, we evaluated our neural network using the evaluation dataset, calculating the mean squared error between the network's predicted value and the dataset's ground truth value. As shown in table 2, we achieved an error of below 0.001 for the circular percussion network, which equates to a standard deviation in size of around 3cm. Whilst for the rectangular percussion network, we achieved an error of $\sim$0.033, and a standard deviation in size of around 20cm. For the rectangular percussion use case, our neural network was penalised by the instruments' spectral relationship between aspect ratio and size, and was not able to generalise to both values as accurately as when size was considered solely on its own. Although the spectral effects of altering the size and aspect ratio of a rectangular percussion instrument should be possible to numerically invert, our network tended to slightly overestimate and then compensate for one of the parameters when performing its predictions. Given that these parameters can be heuristically retrieved independently of one another, this diminishing accuracy is something that we will be intimately conscious of as we begin to extend our model to geometric tasks which are not numerically invertible.

7. DISCUSSION

This work serves as our initial inquiry into the field of orchestration with percussive and inharmonic sounds. So far we have been able to develop a tool that can aid practitioners in determining the size and aspect ratio of circular or rectangular percussion instruments. By approaching orchestration from this technological perspective, we aim to most prominently impact the practices of digital musicians and instrument designers, with the sensibility of composers and orchestrators at the core of our methodology. With such an orchestration tool at our disposal, we may thus be able to extend upon our instinctual aesthetic sensibilities, and curate musical moments that push beyond the boundaries of our situated sonic vocabularies. Using the technologies presented here, it has now become more feasible to determine the material properties of a percussion instrument that may compliment the spectral content of an arbitrary musical fragment. In the wake of these successes and ideas, however, there remains a myriad of ways with which we aim to develop upon this work.

For us, the next step is to extend this work to include geometries beyond just the rudimentary circular and rectangular percussion models that we have so far been working with. By increasing our range of geometries to include many more types of two-dimensional instruments, we aim to extend upon the material details and timbral intricacies that our neural architecture can account for. Some of our work towards this directive has already been completed, having previously published a dataset of convex polygonal drums [36], with more datasets to follow centering around elliptic and concave polygonal percussion instruments. Many of these geometries require individual means of representation, for which we also aim to discover a generalisable means of curating a learned representation of these sonic and geometric objects.

To make sure that our future neural networks develop

a robust relationship between a percussion instrument's timbral signature and its material and semantic descriptions, we also aim to employ and develop our work using a number of additional approaches to percussion synthesis. This will include other approaches to physical modelling, such as the finite difference time domain method, as well as more contemporary approaches based on a neural audio synthesis [40]. By employing these models, and training our future networks on datasets comprised of multiple synthesis methodologies, it is our aim that we may approach a more generalised impression of percussion sounds, attributable to a range of both synthetic and acoustic instruments. It is our impression that a work curated towards a diverse array of synthetic and acoustic soundworlds will increasingly present and refine its own practical and theoretical value. These developments will also involve further situated, practice based evaluations, which will serve to challenge the liminal effectiveness regarding our core understanding of orchestration.

8. CONCLUSION

We have here presented a neural network approach to the parameter estimation of physically modelled, circular and rectangular percussion instruments. Our approach, which centres around a CNN architecture, has enabled us to predict the size and aspect ratio of these percussion instruments to a high degree of accuracy. This neural network serves as our initial attempt towards creating a computer-assisted orchestration methodology specifically targeting the musical employment and arrangement of inharmonic timbres and percussive instruments. Grounded by practice based research and other situated cultural perspectives, we envision these tools being used in a variety of musical settings. The findings and perspectives presented here have highlighted as many questions as they have answered, with plans to continue developing upon and testing within this field in the near future.

9. ACKNOWLEDGEMENTS

The authors would like to thank Francis Devine[3] for his help in preparing figure 1.

This work is supported by the Centre for Doctoral Training in Artificial Intelligence and Music at Queen Mary University of London, funded by UK Research and Innovation (UKRI) under EPSRC grant EP/S022694/1.

[3] Francis Devine: `http://francisdevine.co.uk`

10. REFERENCES

[1] Y. Maresz, "On computer-assisted orchestration," *Contemporary Music Review*, vol. 32, no. 1, pp. 99–109, 2013.

[2] N. Del Mar, *Anatomy of the Orchestra*. London, UK: Faber & Faber Limited, 2nd ed., 1983.

[3] H. Macdonald, *Berlioz's Orchestration Treatise: A Translation and Commentary*. Cambridge, UK: Cambridge University Press, 2002.

[4] S. Adler, *The Study of Orchestration*. New York, NY: W. W. Norton, 4th ed., 2016.

[5] F. Orduña-Bustamante, "Nonuniform beams with harmonically related overtones for use in percussion instruments," *The Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2935–2941, 1991.

[6] F. Soares, J. Antunes, and V. Debut, "Multi-modal tuning of vibrating bars with simplified undercuts using an evolutionary optimization algorithm," *Applied Acoustics*, vol. 173, 2021.

[7] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*. Heidelberg, Germany: Springer, 2005.

[8] W. A. Sethares and W. Vitale, "Ombak and octave stretching in balinese gamelan," *Journal of Mathematics and Music*, vol. 14, pp. 1–17, 2020.

[9] B. Spatz, *What a Body Can Do: Technique as Knowledge, Practice as Research*. Abingdon, UK: Routledge, 2015.

[10] L. Wolstanholme and F. Devine, "terracotta," in *22nd International Conference on New Interfaces for Musical Expression (NIME)*, (Auckland, New Zealand), 2022.

[11] L. Wolstanholme and F. Devine, "josef: Spatiality as a material property of audiovisual art," *International Journal of Creative Media Research (IJCMR)*, vol. Forthcoming, 2023.

[12] L. Wolstanholme and A. McPherson, "Remarks on a cultural investigation of abstract percussion instruments," in *Digital Music Research Network (DMRN+17)*, (London, UK), 2022.

[13] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "Risks from learned optimization in advanced machine learning systems," *arXiv preprint arXiv:1906.01820*, 2019.

[14] M. J. Yee-King, L. Fedden, and M. d'Inverno, "Automatic programming of VST sound synthesizers using deep networks and other techniques," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 150–159, 2018.

[15] O. Barkan, D. Tsiris, O. Katz, and N. Koenigstein, "InverSynth: Deep estimation of synthesizer parameter configurations from audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2385–2396, 2019.

[16] C. Mitcheltree and H. Koike, "SerumRNN: Step by step audio VST effect programming," in *10th International Conference on Artificial Intelligence in Music, Sound, Art and Design*, (Seville, Spain), pp. 218–234, 2021.

[17] D. Sheng and G. Fazekas, "Automatic control of the dynamic range compressor using a regression model and a reference sound," in *20th International Conference on Digital Audio Effects (DAFx)*, (Edinburgh, UK), pp. 160–167, 2017.

[18] J. Rämö and V. Välimäki, "Neural third-octave graphic equalizer," in *22nd International Conference on Digital Audio Effects (DAFx)*, (Birmingham, UK), 2019.

[19] D. Sheng and G. Fazekas, "A feature learning siamese model for intelligent control of the dynamic range compressor," in *International Joint Conference on Neural Networks (IJCNN)*, (Budapest, Hungary), pp. 1–8, 2019.

[20] F. S. Caspe, A. McPherson, and M. Sandler, "DDX7: Differentiable FM synthesis of musical instrument sounds," in *23rd International Society for Music Information Retrieval*, (Bengaluru, India), pp. 608–616, 2022.

[21] L. Gabrielli, S. Tomassetti, S. Squartini, and C. Zinato, "Introducing deep machine learning for parameter estimation in physical modelling," in *20th International Conference on Digital Audio Effects (DAFx)*, (Edinburgh, UK), pp. 11–16, 2017.

[22] H. Han and V. Lostanlen, "WAV2SHAPE: Hearing the shape of a drum machine," in *Forum Acusticum*, (Lyon, France), pp. 647–654, 2020.

[23] G. Carpentier, *Approche computationnelle de l'orchestration musciale - Optimisation multicritère sous contraintes de combinaisons instrumentales dans de grandes banques de sons*. PhD thesis, Université Pierre et Marie Curie, Paris, France, 2008.

[24] G. Carpentier, G. Assayag, and E. Saint-James, "Solving the musical orchestration problem using multiobjective constrained optimization with a genetic local search approach," *Journal of Heuristics*, vol. 16, no. 5, pp. 681–714, 2010.

[25] P. Esling, G. Carpentier, and C. Agon, "Dynamic musical orchestration using genetic algorithms and a spectro-temporal description of musical instruments," in *European Conference on the Applications of Evolutionary Computation*, (Istanbul, Turkey), pp. 371–380, 2010.

[26] M. Caetano and C. E. Cella, "Imitative computer-aided musical orchestration with biologically inspired algorithms," in *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity* (E. R. Miranda, ed.), pp. 585–615, Cham, Switzerland: Springer International Publishing, 2021.

[27] C. E. Cella, "Orchidea: A comprehensive framework for target-based computer-assisted dynamic orchestration," *Journal of New Music Research*, pp. 1–29, 2022.

[28] A. Fraser and D. Burnell, *Computer Models in Genetics*. New York, NY: McGraw-Hill Inc., 1970.

[29] J. Gillick, C. E. Cella, and D. Bamman, "Estimating unobserved audio features for target-based orchestration," in *20th International Society for Music Information Retrieval Conference (ISMIR)*, (Delft, The Netherlands), pp. 192–199, 2019.

[30] C. E. Cella, L. Dzwonczyk, A. Saldarriaga-Fuertes, H. Liu, and H.-C. Crayencour, "A study on neural models for target-based computer-assisted musical orchestration," in *2020 Joint Conference on AI Music Creativity*, (Stockholm, Sweden), 2020.

[31] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CRePE: A convolutional representation for pitch estimation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Calgary, Canada), pp. 161–165, 2018.

[32] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and musica)," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[33] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *39th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Florence, Italy), pp. 659–663, 2014.

[34] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations (ICLR)*, (San Diego, CA), 2015.

[36] L. Wolstanholme, "kac_drumset: A dataset generator for arbitrarily shaped drums." Zenodo, 2022. Version: 1.1.

[37] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. New York, NY: Springer, 2nd ed., 1998.

[38] S. Bilbao, *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*. Chichester, UK: John Wiley & Sons, 2009.

[39] N. H. Asmar, *Partial Differential Equations and Boundary Value Problems with Fourier Series*. Upper Saddle River, NJ: Pearson Prentice Hall, 2nd ed., 2004.

[40] R. Diaz, B. Hayes, C. Saitis, G. Fazekas, and M. Sandler, "Rigid-Body Sound Synthesis with Differentiable Modal Resonators," in *48th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Rhodes Island, Greece), 2023.