Towards multisensory control of physical modeling synthesis

Loïc Jankowiak, Han Han, Vincent Lostanlen, Mathieu Lagrange

25·29 August 2024
inter.noise
NANTES FRANCE

Towards multisensory control of physical modeling synthesis

Loïc Jankowiak
Independent researcher
44000 Nantes, France

Han Han
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004
1, rue de la Noë, 44000 Nantes, France

Vincent Lostanlen
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004
1, rue de la Noë, 44000 Nantes, France

Mathieu Lagrange
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004
1, rue de la Noë, 44000 Nantes, France

ABSTRACT

Physical models of musical instruments offer an interesting tradeoff between computational efficiency and perceptual fidelity. Yet, they depend on a multidimensional space of user-defined parameters whose exploration by trial and error is impractical. Our article addresses this issue by combining two ideas: query by example and gestural control. On one hand, we train a deep neural network to identify the resonator parameters of a percussion synthesizer from a single audio example via an original method named perceptual—neural—physical sound matching (PNP). On the other hand, we map these parameters to knobs in a digital controller and configure a musical touchpad with MIDI polyphonic expression. Hence, we propose a multisensory interface between human and machine: it integrates haptic and sonic information and produces new sounds in real time as well as visual feedback on the percussive touchpad. We demonstrate the interest of this new kind of multisensory control via a musical game in which participants collaborate with the machine in order to imitate the sound of an unknown percussive instrument as quickly as possible. Our findings show the challenge and promise of future research in musical "Human–AI parternships".

1. INTRODUCTION

Can a musical instrument respond to multiple senses at once? Over the past four decades, the development of machine listening technologies have shown the value of integrating auditory interactions into cyber-human musicianship [1]. For example, machine listening in Antescofo allows a soloist to control the pace of computer-generated accompaniment in real time, simply by playing through a microphone [2]. Another example is found in Orchidea, a software framework for target-based automatic orchestration [3]. More recently, the advent of neural audio models has paved the way towards end-to-end generation of polyphonic music from a live audio input [4].

Meanwhile, the research community on new interfaces for musical expression (NIME) has built custom controllers for sound generation [5]. Pressure-sensing surfaces offer a successful example of such controllers: indeed, their layouts are reminiscent of drum membranes, making them intuitive to percussionists [6]. Compared to the conventional setup of mouse and keyboard, touch surfaces have numerous advantages for live musical performance: finer control of stroke position and velocity, offering a more embodied sense of rhythm and easier visual coordination with other musicians [7].

Yet, compared to the extensive prior work on unisensory control—i.e., either auditory *or* tactile—the important question of multisensory control—both auditory *and* tactile—remains insufficiently explored. This gap in knowledge is partly attributable to a disconnect between music information retrieval (MIR) and contemporary musicianship [8]. In addition, the collection of human judgments for interaction design is costly and time-consuming: thus, supervised machine listening models must be trained on synthetic data or surrogate tasks. Lastly, general-purpose "foundation models" for audio (e.g., MERT [9]) are certainly effective for MIR tasks but remain difficult to interpret for humans. This raises the long-standing issue of coming up with an intuitive mapping for timbre as a musical control structure [10].

In this paper, we propose a coherent framework for multisensory control of musical sounds, integrating auditory, haptic, and visual sources of information. For this purpose, we build upon well-established knowledge on physical modeling synthesis. Compared to waveform-based neural audio synthesizers, physical models have the advantage of being directly interpretable in terms of sound production. This interpretation allows interaction design with enhanced iconicity; i.e., the cognitive analogy between human gesture and machine response. However, the exploration of musical timbre should not be solely be guided by physics but also by auditory perception, as Risset famously advocated [11]. Hence, we supplement the physical synthesis model by a state-of-the-art machine listening model which is trained to recover synthesizer controls from audio according to a perceptually plausible learning objective. In practice, the machine listening prediction is judged to be imperfect. Yet, since it is interoperable with the synthesizer, it may serve as a preset which is learned on the fly and fine-tuned by humans.

Loosely speaking, human and machine "listen together" to a given sound and "search together" for a matching synthetic sample. More precisely, they build a kind of human–computer partnership [12]. An important prior publication on this topic is [13], who developed acoustical and behavioral search heuristics for interactive sound design of electric vehicles. In comparison, the originalities of our work are: physical modeling synthesis; gestural mapping with iconicity; perceptual–neural–physical sound matching (PNP); and the collection of real-time behavioral data beyond mere judgments of pleasantness.

2. METHODS

2.1. Physical Model

The vertical displacement $\boldsymbol{x}$ of a rectangular drum membrane in the Cartesian coordinate system $\boldsymbol{u} = \{u_1, u_2\}$ can be described by a fourth-order partial differential wave equation in dimension two. Namely, for $t \geq 0$:

$$\left(\frac{\partial^2 \boldsymbol{x}}{\partial t^2}(t,\boldsymbol{u}) - c^2\nabla^2\boldsymbol{x}(t,\boldsymbol{u})\right) + S^4\left(\nabla^4\boldsymbol{x}(t,\boldsymbol{u})\right) + \frac{\partial}{\partial t}\left(d_1\boldsymbol{x}(t,\boldsymbol{u}) + d_3\nabla^2\boldsymbol{x}(t,\boldsymbol{u})\right) = 0$$

S, c, d_1, d_3 are respectively material stiffness, traveling sound speed through the membrane, frequency-independent and frequency-dependent damping coefficients. The membrane has a length of l and width of $l\alpha$, $\alpha \in (0,1]$. It is bounded at zero at all time and set to vibrate when releasing at an initial vertical displacement of h at a given location $\boldsymbol{u}_0$.

We adopt the well-established Functional transformation method (FTM) [14] to solve the above PDE. FTM transforms the PDE into subsequently Laplace and Sturm-Liouville domains to

Figure 1: Left: screenshot of the synthesizer plugin. Right: Experimental setup of the sound matching game.

derive an algebraic solution of SLT $\circ\, \mathscr{L}(\boldsymbol{x})(s,\mu)$ in the functional spaces. Then, it applies inverse Laplace transform and inverse Sturm-Liouville transform to obtain a solution in space–time. The solution follows a modal synthesis form

$$\boldsymbol{x}(t,\boldsymbol{u}) = \sum_{\boldsymbol{m}\in\mathbb{N}^2} K_m(\boldsymbol{u},t)\exp(\sigma_{\boldsymbol{m}}t)\sin(\omega_{\boldsymbol{m}}t), \tag{1}$$

with modal frequencies $\omega_{\mathbf{m}}$, modal amplitudes K_m and modal exponential decay rates σ_m. The modal coefficients are

$$\omega_{\mathbf{m}}^2 = \left(S^4 - \frac{d_3^2}{4}\right)\Gamma_{\mathbf{m}}^2 + \left(c^2 + \frac{d_1 d_3}{2}\right)\Gamma_{\mathbf{m}} - \frac{d_1^2}{4} \tag{2}$$

$$\sigma_{\mathbf{m}} = \frac{d_3}{2}\Gamma_{\mathbf{m}} - \frac{d_1}{2} \tag{3}$$

$$K_{\mathbf{m}}(\boldsymbol{u},t) = y_u^m \delta(t)\sin\left(\frac{\pi m_1 u_1}{l}\right)\sin\left(\frac{\pi m_2 u_2}{l\alpha}\right) \tag{4}$$

where $\Gamma_{\mathbf{m}} = \pi^2 m_1^2/l^2 + \pi^2 m_2^2/(l\alpha)^2$, and $y_u^{\mathbf{m}}$ is the $\mathbf{m}^{th}$ coefficient associated to the eigenfunction $\sin(\pi\mathbf{m}\boldsymbol{u}/l)$ that decomposes $y_u(\boldsymbol{u})$.

To offer more intuitive control, we reparametrize the PDE parameters $\{S, c, d_1, d_3, \alpha\}$ into $\{\omega_1, \tau_1, p, D, \alpha\}$, which better informs the perceptual quality of the sounds. This conversion is detailed in [15], where $\{\omega_1, \tau_1, p, D, \alpha\}$ correspond to respectively fundamental frequency, duration, inhomogeneous damping, inharmonic dispersion and aspect ratio of the membrane.

We implement the above physical modeling synthesis algorithm in a C++ plugin using JUCE library. The compiled format is compatible with standard digital audio workstations. The plugin, as shown in Figure 1 (left), maps the received MIDI values of the 5 knobs labeled "Pitch", "Sustain", "Damp", "Inharmonicity", and "Squareness", to the control parameters $\{\omega_1, \tau_1, p, D, \alpha\}$, respectively. When a MIDI note is triggered, its velocity is mapped to the initial displacement h. An additional visualization panel is provided to see the shape of and the excitation position on the membrane. Polyphonic percussive sound is rendered in real time, given the input parameters and excitation positions (X, Y). The plugin is open source and available on Github [1].

2.2. Machine listening

Sound matching involves distinguishing spectrotemporal characteristics and identifying the controls instrumental to narrowing the perceived difference. Such a task may be significantly eased with machine listening. In [16], several EfficientNet-B0 type convolutional neural networks are trained with various loss functions to predict the synthesis input to the FTM drum synthesizer of a given target sound. The dataset comprises 100k synthesized sounds in total, with a

[1] FTM Synth Github repository: https://github.com/Synthesis/FTMSynth

train/test/validation split of 8 : 1 : 1. Fundamental frequency ω_1 ranges between 40 Hz and 1 kHz; duration τ_1, between 400 ms and 3 s; inhomogeneous damping rate p, between 10^{-5} and 0.2; frequential dispersion D, between 10^{-5} and 0.3; and aspect ratio α, between 10^{-5} and 1.

Using two perceptually relevant spectral representations as evaluation metrics: the L^2 distance of Joint time–frequency scattering coefficients and multiscale spectrogram loss, the best accuracy was achieved via pretraining with parameter loss and finetuning with perceptual–neural–physical (PNP) loss. To incorporate machine listening into the sound matching pipeline, we select two model checkpoints: the best-performing PNP model and the previous state-of-the-art model trained with parameter loss. For a given target sound, we extract its constant-Q transform (CQT) coefficients as done in [16] and apply one forward pass through the loaded checkpoint models. The sound matching output predicted by the above two models will be used to initialize the probing synthesizer in the subsequent sound matching experiments.

2.3. Pressure-sensitive controller

The Erae Touch, manufactured by Embodme, is a rectangular pressure-sensitive controller for MIDI Polyphonic Expression (MPE)[2]. Its 18-inch silicone touch surface allows for detecting expressive gestural control at precise x/y locations. An LED matrix under the pressure-sensitive surface serves as a display for visual feedback, and the device's layout can be fully customized using the Erae Lab software.

For the needs of this experiment, we are using a single drum pad covering the entire surface of the device. The location of the user's input on the drum pad is sent to the plugin via MIDI, and mapped to the excitation position (X, Y) in our physical model.

2.4. Sound Matching Experiments

Our corpus comprises five synthetic percussive sounds generated by the physical model at position $(0.4l, 0.4l\alpha)$. They are selected at random to have varied spectral compositions evoking different materials. Our goal is to observe how human subjects match these given target sounds with and without the aid of a machine listening model.

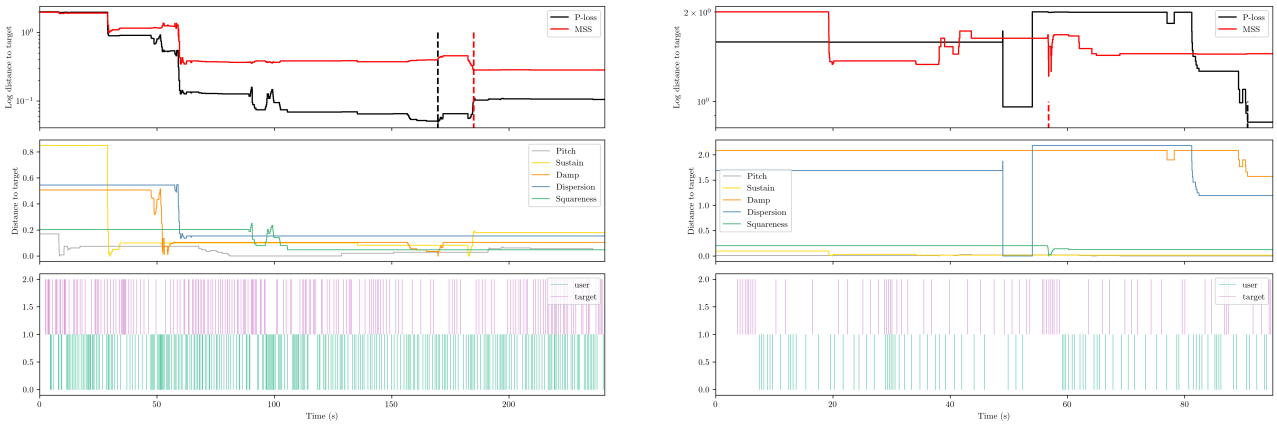
We recruit 16 participants of all genders, aged 23–59, with varied amounts of musical experience. The experimental setup consists of two Erae Touch drum pads and a MIDI controller (Behringer X-Touch Mini) in Fig. 1 (right).

The target sound is programmed onto the right-hand Erae Touch and is unalterable. The participant is asked to adjust five knobs on the MIDI controller controlling the left-hand Erae Touch, such that its sound approaches as close as possible to the target. We present the five sounds in the same order to all participants. We initialize the left-hand Erae Touch to a drum either at random, or at the predicted output of a neural network trained with PNP loss or parameter loss. Note that despite being a single sound, each target implies a unique physical drum capable of eliciting a far bigger variety of sounds on the full touch surface.

At each session, the participant is introduced to their task and explained the functionalities of each control knob. They have a few minutes of "practice time" to freely explore the synthesizer before starting the game. For each sound, the participants are given 4 minutes as the maximum time limit, during which they are free to explore the entire drum surface and find the optimal matching parameter set. The participants are also allowed to stop early if they do not think any improvement can be made.

We record the matching trajectories of each participants, including the MIDI notes (note on/note off) and CC (Control Change) events triggered over the course of four minutes.

[2]Official website: https://www.embodme.com/erae-touch



(a) Random initialization

(b) PNP-predicted initialization

Figure 2: Example trajectories of a participant matching the target sound from a given initialization. The two vertical dashed lines in bright red and black are the global minima of multiscale spectral loss (MSS) and parameter loss (P-loss), respectively. The purple and green vertical lines are the target and probe Erae Touch hits by the participants. PNP stands for perceptual–neural–physical sound matching.

3. RESULTS AND DISCUSSIONS

3.1. Hypothesis

Given the high accuracy of sound matching pre-trained models in Section 2.2, we hypothesize that machine listening-aided sound matching reaches higher accuracy than matching from random.

3.2. Evaluation

We use individual parameter loss, mean squared error of parameters' losses (P-loss), and multiscale spectrogram loss (MSS) [17] as evaluation metrics of the sound matching accuracy. Multi-scale spectrogram loss is the added logarithmic and linear L1 distance of spectrograms computed with various window sizes. We adopt the implementation in [18] and its default settings. While P-loss evaluates the precise parametric error, MSS implies perceived perceptual difference. The two metrics are correlated in general but may differ in certain parameter regions. For example, two sounds with overlapping partials but different fundamental frequencies may have a much more significant P-loss than MSS loss. As indicated in the upper subplots of Fig. 2, the minimum accuracy achieved by P-loss and MSS loss do not always coincide at the same time stamp.

We choose to use the matched sound with minimal error to compare performance across trials and participants. In general, it is difficult to measure one's performance when no clear indication of the "finalized satisfactory answer" is provided by the participant. Extracting the last matched sound in time overlooks the fact that a more similar sound may have been attempted before. Taking into account only the sound with minimal error however, risks concealing the diverging trajectory afterwards. Considering the difficulty of our task and the frequent occurrences of participants diverging from the initialized sound given by machine listening models, we consider minimal error an overall better indicator of each trial's performance.

3.3. Results and discussions

We report minimum parameter loss (P-loss) and multi-scale spectrogram loss (MSS) for all matching trajectories in Fig. 3, accounting for 16 participants and 3 initialization methods.

Despite having varying expertise, all participants are able to improve upon the initialized sound. The overall trend of the minimum accuracy can be summarized as follows: for target

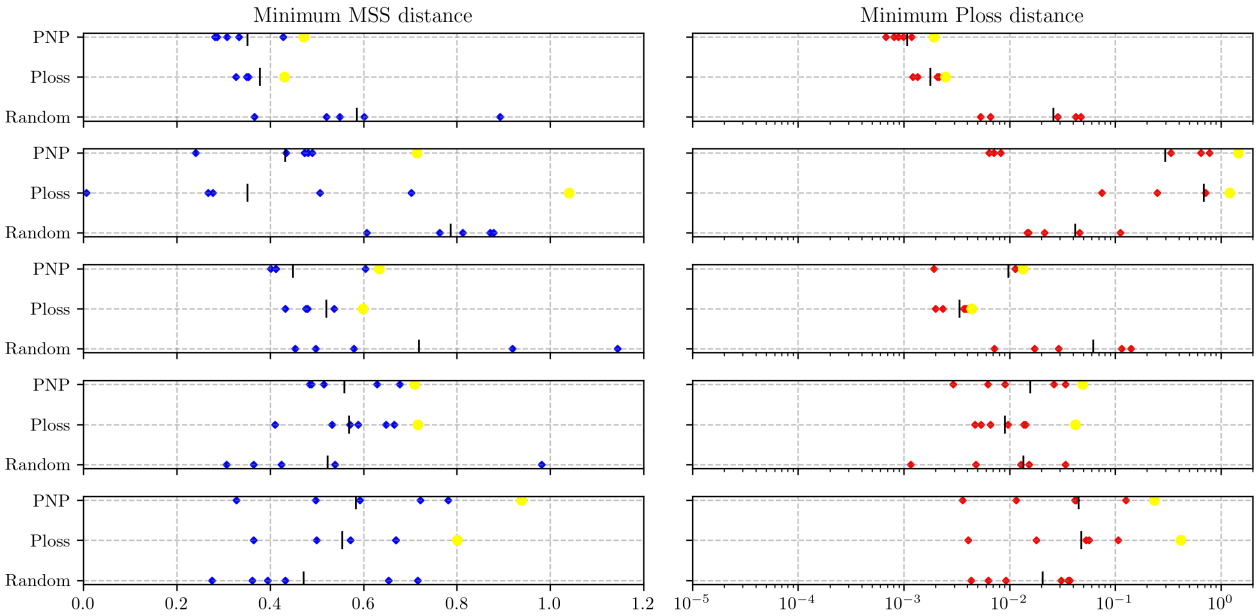


Figure 3: Minimum sound matching error achieved with each of the five sounds in the corpus. Each row corresponds to a different sound. In blue is minimum Multi-scale spectrogram loss (MSS) of different trials, in red is minimum parameter loss of different trials. Black vertical lines indicate the mean value across trials. Yellow circles indicate the MSS and Ploss distance at initialization.

sound 1 and 3, machine listening models improve sound matching accuracy on both P-loss and MSS metrics. For target sound 2, machine listening models improve perceived difference but deteriorate P-loss. For targets 4 and 5, machine listening models do not have an effect on the minimum achieved accuracy.

For a given target sound and initialization setup, the performance across trials performed by different participants may vary a lot. For example the initialization at random trials of target sound 3 have a significant standard deviation, whereas the trials corresponding to starting from PNP model prediction for target sound 1 has very small variance in performance. These discrepancies in performance may result from either expertise, complexity of the sounds themselves, fatigue as the experiment proceeds, among other factors. We find the sound matching performance overall unrelated to years of musical practice. Failing to understand the control knobs associated to high level description of the partials distribution such as dispersion and damping, certain participants may enter a phase of frustration and assert random attempts to fill the time limit, rendering the data irrelevant. Certain knobs also have granularities that do not vary linearly with perception. In these circumstances, it is easy to diverge suddenly and hard to return. Overally speaking, the difficult nature of this task posed complications in evaluating the performance consistently.

4. CONCLUSION

In this paper, we have combined physical modeling synthesis, gestural mapping, machine listening with neural networks, and visual feedback of haptic interactions. The overarching goal behind this framework is to better understand the synergy of multiple senses (hearing, touch, vision) in the creative work of sound designers and electronic music artists. For this purpose, we have implemented a 2D rectangular drum physical model in C++ and integrated it with a pressure-sensitive rectangular controller. With this integration, the framework produces both high-quality audio data and high-quality gestural data at an affordable cost. We have recruited 16 participants to partner with the machine in sound matching tasks on a corpus of five sounds.

We have compared the influence of initialization: i.e., either random or informed by the predicted output of two machine listening models. Our main finding is that human approaches to sound matching are highly varied and cannot be reduced to be one single strategy. At the moment, our hypothesis seems to hold for certain sounds but not for others; there is no strong evidence of machine listening aiding the convergence of sound matching in general. Future work is needed to analyze the acquired multisensory data in finer detail and understand how machine listening can become a reliable feature in pressure-sensitive touch surfaces. In particular, the machine listening component in our framework informs the human only at the initial stage; the question of designing *sustained* multisensory interactions for time-efficient sound design remains open.

ACKNOWLEDGEMENTS

We thank Thomas Canal and Timothée Mesnard, for their contribution to the pilot study; Khrystyna Povkh and Emily Thureau, for administrative support; Adeline Hérau, for vocational counseling; and all the participants to the study. V. Lostanlen is supported by ANR project nIrVAna (ANR-23-CE37-0025-04).

REFERENCES

1. Robert Rowe. *Interactive music systems: Machine listening and composing*. MIT press, 1992.
2. Arshia Cont. Antescofo: Anticipatory synchronization and control of interactive parameters in computer music. In *International Computer Music Conference (ICMC)*, pages 33–40, 2008.
3. Carmine-Emanuele Cella. Orchidea: a comprehensive framework for target-based computer-assisted dynamic orchestration. *Journal of New Music Research*, 51(1):40–68, 2022.
4. Théo Jourdan and Baptiste Caramiaux. Machine learning for musical expression: A systematic literature review. In *New Interfaces for Musical Expression (NIME)*, 2023.
5. Alexander Refsum Jensenius and Michael J. Lyons. *A NIME reader: Fifteen years of new interfaces for musical expression*, volume 3. Springer, 2017.
6. Diemo Schwarz, Wanyu Liu, and Frédéric Bevilacqua. A survey on the use of 2d touch interfaces for musical expression. In *New Interfaces for Musical Expression (NIME)*, 2020.
7. Anna Xambó. Embodied music interaction: creative design synergies between music performance and hci. *Digital Bodies: Creativity and Technology in the Arts and Humanities*, pages 207–220, 2017.
8. Carmine-Emanuele Cella. Music information retrieval and contemporary classical music: A successful failure. *Trans. Int. Soc. Music. Inf. Retr.*, 3(1):126–136, 2020.
9. Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, et al. MERT: Acoustic music understanding model with large-scale self-supervised training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
10. David L. Wessel. Timbre space as a musical control structure. *Computer music journal*, pages 45–52, 1979.
11. Jean-Claude Risset and David L Wessel. Exploration of timbre by analysis and synthesis. In *The psychology of music*, pages 113–169. Elsevier, 1999.
12. Wendy E Mackay. Creating human-computer partnerships. In *International Conference on Computer-Human Interaction Research and Applications*, pages 3–17. Springer, 2023.
13. Ava Souaille, Vincent Lostanlen, Mathieu Lagrange, Nicolas Misdariis, and Jean-François Petiot. Acoustical and behavioral heuristics for fast interactive sound design. *Plos one*, 19(1):e0296347, 2024.
14. L. Trautmann and Rudolf Rabenstein. *Digital Sound Synthesis by Physical Modeling Using the Functional Transformation Method*. Springer, 2003.

15. Han Han and Vincent Lostanlen. wav2shape: Hearing the Shape of a Drum Machine. In *Proceedings of Forum Acusticum*, 2020.
16. Han Han, Vincent Lostanlen, and Mathieu Lagrange. Perceptual–Neural–Physical Sound Matching. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, June 2023.
17. Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable Digital Signal Processing. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
18. Christian J. Steinmetz and Joshua D. Reiss. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020.