# *Melia*: An Expressive Harmonizer at the Limits of AI

Matthew Caren
Joshua Bennett
{mcaren,joshuab}@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

## Abstract

We present *Melia*, a digital harmonizer instrument that explores how common failure modes of machine learning and artificial intelligence (ML/AI) systems can be used in expressive and musical ways. The instrument is anchored by an audio-to-audio neural network trained on a hand-curated dataset to perform pitch-shifting and dynamic filtering. Biased training data and poor out-of-distribution generalization are deliberately leveraged as musical devices and sources of instrument-defining idiosyncrasies. *Melia* features a custom hardware interface with a MIDI keyboard that polyphonically allocates instances of the model to harmonize live audio input, as well as controls that manipulate model parameters and various audio effects in real-time. This paper presents an overview of related work, the instrument itself, and a discussion of how audio-to-audio AI models might fit into the long-standing tradition of musicians, artists, and instrument-makers finding inspiration in a medium's shortcomings.

## Keywords

Harmonizer, AI, Failure, Voice, NIME

## 1 Introduction

Artists have long found inspiration in the limitations of technology's ability to capture and understand the world. Artist Brian Eno noted that: *"whatever you now find weird, ugly, uncomfortable and nasty about a new medium will surely become its signature. CD distortion, the jitteriness of digital video, the crap sound of 8-bit...it's the sound of failure...of a medium pushing to its limits and breaking apart"* [6]. Further, that when technology *"fails conspicuously, and especially if it fails in new ways, the listener believes something is happening beyond its limits."*

We present *Melia*, an instrument built around an audio-to-audio AI model pushed to fail—trained on limited, biased training data and performed in adversarially noisy outdoor environments—in an exploration of how the failure modes of this new technology might be employed as expressive, musical tools. At its core, the instrument is a harmonizer: a musician sings into a microphone and uses a piano keyboard interface to polyphonically manipulate synthetic copies of their voice. Harmonizers typically use non-ML digital signal processing (DSP) techniques to perform this task [2, 3, 7, 11]—with *Melia*, we are not aiming to outperform existing harmonizers, but rather to use the affordances and expressivity of harmonizer-type instruments to explore new interactions and sonic paradigms in the failure modes of AI models.

The NIME community has long been interested in both ML/AI [8] and the human voice [9] as gateways to new sounds and interfaces. Utilizing failure in NIMEs as a means for musical expression has been explored in the sense of *human* failure [10], but less explicitly for *technological* failure. That said, the idea of building instruments around the idiosyncrasies of how technologies "fail" is certainly not new: examples include distortion from vacuum tube guitar amplifiers, saturation of Moog-style ladder filters, and wow and flutter in magnetic tape-based delay and modulation effects. Many extended techniques can also be considered controlled, expressive uses of "failures" of instruments' intended playing modes, from overblowing on woodwinds to electric guitar feedback.

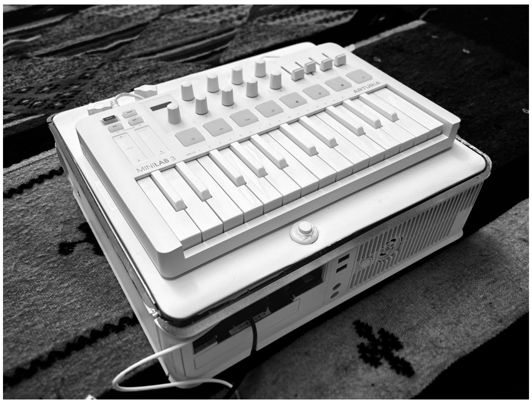

**Figure 1: The *Melia* instrument**

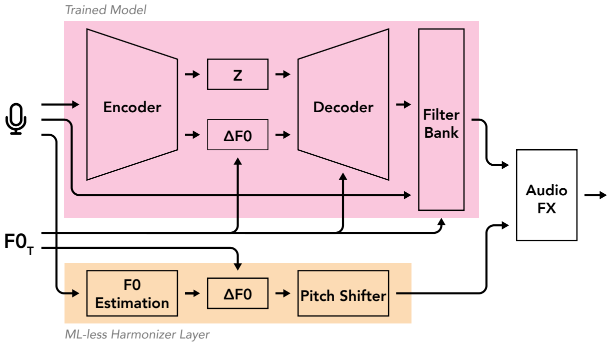

**Figure 2: Diagram of the signal path of *Melia***

## 2 *Melia*

An overview of *Melia*'s signal processing chain is given in Figure 2. This signal path is duplicated for each voice, enabling up to 8-voice polyphony.

### 2.1 Model

The *Melia* model is a convolutional autoencoder trained to perform pitch shifting and dynamic filtering. The encoder extracts

*NIME '25, June 24–27, 2025, Canberra, Australia*

an estimated pitch $F0_E$ and a condensed timbre representation $Z$ from the input audio. The target pitch of the voice $F0_T$ is used to calculate the required relative frequency shift $\Delta F0 = F0_T - F0_E$, which is passed to the decoder along with the absolute target pitch $F0_T$ and timbral representation $Z$ to be processed into a shifted result. The signal is then mixed with the original input audio and passed through a final bank of parameterized, trainable feedback comb filters.

The model is trained in a two-step process: first, we train just the $F0_E$ extractor. Then, we freeze the $F0_E$ extractor weights and train the rest of the model end-to-end using pairs of samples generated by the same synthesizer at different pitches using multi-scale spectrogram (MSS) loss.

The training data is taken from a small dataset of audio clips generated using a variety of simple subtractive synthesizer patches [4]. By carefully curating the training dataset, we induce symptoms of two of the most common failure modes in AI/ML systems: *out-of-distribution generalization* and *training data bias*. The model is parameterized with a *tonic* frequency, which is disproportionately favored as a target output pitch. In the trained model, we observe that this causes several of the comb filters to emphasize the tonic as a resonant frequency—with sufficiently broadband input, this sets up a drone at the tonic frequency and its corresponding harmonics.

Because the training data has virtually no noise, the performance of the trained model is highly vulnerable to noisy input audio. When *Melia* is performed outdoors, the user must sing loudly to overcome the high noise floor. When the user sings quietly, environmental noise interferes with the model's ability to track and modify pitch. When a singing voice is not present at all, the model continues to search for a stable tone to tune—which it may occasionally find in birdsong or wind howling—and this oscillation between occasional success and repeated failure creates a sonically rich and unpredictable texture.

This setup allows a musician to play with the contrast between the "in-distribution" inputs of a pitched singing voice and the "out-of-distribution" inputs of complex environmental noise.

### 2.2 ML-less Harmonizer Backbone

Unpredictable black boxes can be difficult and frustrating interfaces for artists trying to realize an artistic vision [1]. Because the model is intentionally designed *not* to be robust, we wanted to be careful that the instability or unpredictability of the output would not inhibit the clarity of the pitches a musician plays. To this end, the model output is supplemented with a "conventional" polyphonic harmonizer backbone, which has a far more stable output. This non-ML harmonizer layer can be mixed with the model output audio using a hardware fader on the instrument.

Each voice of the harmonizer backbone is synthesized by estimating the F0 of the input using the YIN algorithm [5] and pitch-shifting it by the difference between the input's intended pitch and the input's estimated pitch. Source code for the non-ML harmonizer layer is open source and publicly available.[1]

### 2.3 Hardware

*Melia* is built into the enclosure of a recycled tower computer—we remove the internal PCBs, rewire several of the internal lights for custom control, and reroute the USB ports to connect to the MIDI keyboard, microphones, and hardware buttons and sliders. Two microphones are mounted inside the instrument, one directed towards the user to capture singing and the other directed away to capture ambient sound. A Teensy microcontroller handles signal routing, lighting, and A/D conversion, while an offboard laptop computer linked via several parallel USB connections is used for model inference and real-time signal processing.

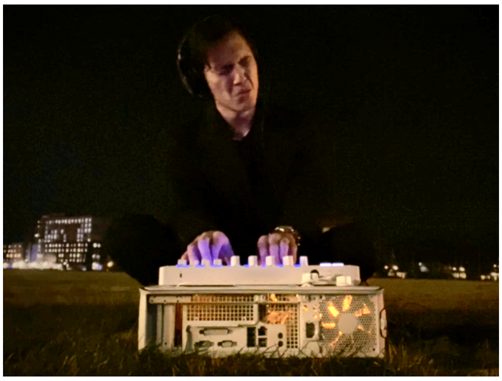

**Figure 3: A live performance of *Melia* in an outdoor field**

Hardware knobs and sliders control key instrument parameters, including the mix between the two microphones and a variety of real-time effects: a feedback delay network reverb, granular delay, and peaking equalizer filter section.

Taking inspiration from the harmonizer built by Ben Bloomberg for Jacob Collier [3], a "freeze" function also allows the user to infinitely sustain voices. This is controlled by an arcade-style momentary button placed directly in front of the keyboard, which allows a user to hold the button with their thumb while keeping their other fingers available to play notes on the keyboard.

## 3 Discussion & Future Work

Due to the constraints of FFT window sizes and model inference times, the instrument has significant end-to-end latency (up to 80 ms). Though this is not enough to break the impression of a live performance from a listener's perspective, it can make performance challenging for a musician. In live settings, we circumvent this issue by mixing *Melia*'s output with an off-the-shelf hardware vocoder as an approximate but low-latency monitor.

In future work, we hope to migrate to a Field Programmable Gate Array (FPGA) to speed up model inference, and more generally to further explore how deliberate biases in training data can be used to induce desired musical effects.

## 4 Ethical Standards

No conflicts of interest were identified. All training data was hand-created for this project by the researchers.

## Acknowledgments

This research was supported by the Council for the Arts at MIT.

## References

[1] Maneesh Agrawala. 2023. Unpredictable black boxes are terrible interfaces.
[2] Antares. 2025. Harmony Engine.
[3] Benjamin Bloomberg. 2020. *Making Musical Magic Live.* Ph. D. Dissertation. Massachusetts Institute of Technology.
[4] Matthew Caren. 2025. *Six-hour Subtractive Synth Sample Set (S5).* https://doi.org/10.5281/zenodo.15233533
[5] Alain De Cheveigné and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111, 4 (2002).
[6] Brian Eno. 2020. *A year with swollen appendices: Brian Eno's Diary.* Faber & Faber.

[1] https://github.com/matthewcaren/tiny-harmonizer

[7] Juan Gremes, Nicola Palavecino, Lucas Seeber, and Santiago Herrero. 2015. Synthetic Voice Harmonization: A Fast and Precise Method. In *2015 IEEE International Symposium on Multimedia (ISM)*.

[8] Théo Jourdan and Baptiste Caramiaux. 2023. Machine Learning for Musical Expression: A Systematic Literature Review. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Mexico City, Mexico.

[9] Rébecca Kleinberger, Nikhil Singh, Xiao Xiao, and Akito van Troyer. 2022. Voice at NIME: a Taxonomy of New Interfaces for Vocal Musical Expression. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. The University of Auckland, New Zealand.

[10] Zeynep Özcan and Anıl Çamcı. 2024. Juggling for Beginners: Embracing and Fabricating Failure as Musical Expression. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 138–141.

[11] TC-Helicon. 2025. Voicelive Series.