

# A MULTI-RESOLUTION SPECTROGRAM APPROACH FOR ESTIMATING THE PHYSICAL PARAMETERS OF A PLATE REVERB

Jared Lipkin, Meiyang Chen, Benjamin R. Thompson, Andrea Cogliati and Michael C. Heilemann

Dept. of Electrical and Computer Engineering  
University of Rochester  
Rochester, NY  
mheilema@ur.rochester.edu

David A. Anderson

Dept. of Electrical Engineering  
University of Minnesota Duluth  
Duluth, MN

## ABSTRACT

The ResNet-18 image classification model is employed to determine the physical parameters of a plate reverb from a recording of the impulse response. The model is adapted to derive parameters using normalized and down-sampled multi-resolution spectrograms computed from the provided impulse responses (IRs). To refine the prediction of the output location, the spectral phase response is also included as an additional input channel to the network since multiple output locations can give the same magnitude response for high-order resonant modes. On a 5000 IR validation set, our model achieves an average normalized mean squared error (NMSE) of 0.02920 across all parameters, with the lowest average NMSE occurring for parameters  $y_o$  (0.00228),  $L_y$  (0.00347), and  $x_o$  (0.00574).

## 1. INTRODUCTION

Task A is focused on identifying six derived parameters of a plate reverb based on a recorded impulse response. These parameters are derived from the physical parameters of the plate, which include density  $\rho$ , thickness  $h$ , Young's modulus  $E$ , and tension  $T_0$ , given constant values for the plate width  $L_x$ , input location  $(x_i, y_i)$ , decay times at DC and 500 Hz  $\tau_0$  and  $\tau_1$ , and Poisson's ratio  $\nu$ . The derived parameters to be identified are density-thickness ratio

$$\mu = \rho h, \quad (1)$$

rigidity ratio  $D/\mu$  where,

$$D = \frac{Eh^3}{12(1-\nu^2)}, \quad (2)$$

tension ratio  $T_0/\mu$ , plate height  $L_x$  and output location  $(x_o, y_o)$ .

Our solution employs a modified version of ResNet-18, which is an image classification model featuring 18 convolution layers and roughly 11.7 million parameters by default [1]. In this application, the model is trained using audio data by first transforming the data into two-dimensional spectrograms.

## 2. TASK-SPECIFIC MODIFICATIONS

Our solution makes several small adaptations to ResNet-18. First, the default starting weights were bypassed and the output layer was

modified to target the seven raw parameters ( $L_y, h, \rho, T_0, E, x_o$  and  $y_o$ ). Also, in the standard usage of ResNet-18, there are three 224x224 channels containing R, G, and B color information. Our usage instead inputs three spectrograms of multiple FFT lengths  $[N = 256, 512, 1024]^1$  such that the input features comprise a multi-resolution spectrogram. An additional fourth channel also inputs spectral phase information to the model with  $N = 512$ . Spectrogram stride length is set at  $N/8$ . From [2, 3] the vibration response of a plate may be modeled as a superposition of resonant modes. For each mode, there are multiple output locations on the plate that may give the same magnitude response due to symmetry between the nodes and anti-nodes of each plate mode. This is especially true for high-order modes that have many nodal lines. For example, normalized output locations  $(x_o, y_o)$  of (0.25, 0.5) and (0.75, 0.5) would give the same magnitude response for the (2,1) mode, but could be distinguished by looking at the phase response relative to other modes. An example of the 4-channel input image is shown in Fig. 1.

### 2.1. Input Pre-Processing

The resonant frequencies of a simply-supported plate as described in the companion document for the challenge problem are given by

$$\omega_m = \frac{\gamma_m}{\sqrt{\rho h}} \sqrt{T_0 + D\gamma_m^2}, \quad (3)$$

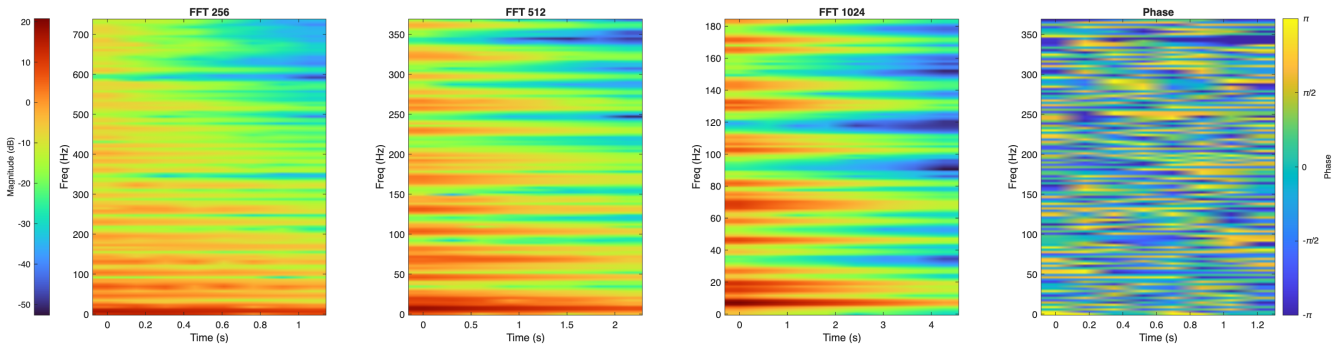
where

$$\gamma_m = \sqrt{\left(\frac{m_x \pi}{L_x}\right)^2 + \left(\frac{m_y \pi}{L_y}\right)^2} \quad (4)$$

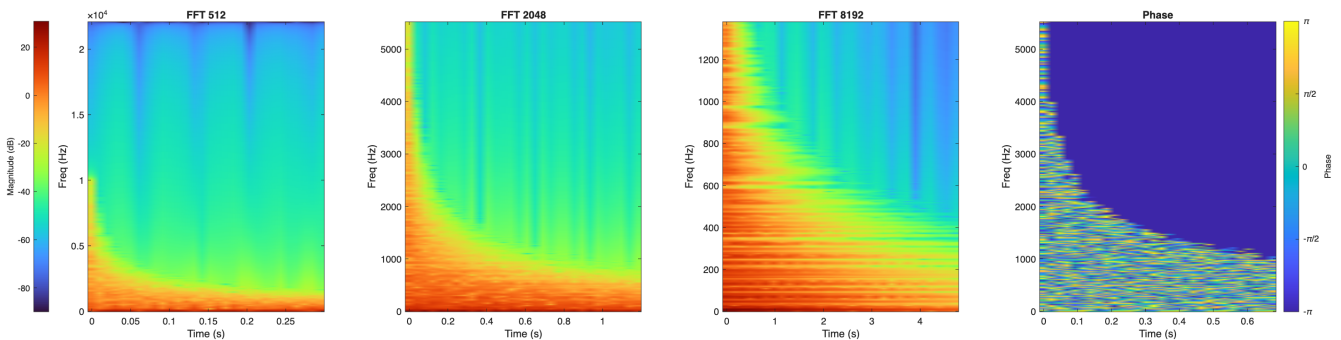
and  $m$  is the mode number with  $m_x$  and  $m_y$  representing the number of half-sinusoids respectively in the horizontal and vertical dimensions.

From [4] the plate response may be separated into two frequency regions based on modal overlap ( $MO$ ), a quantity that describes the number of modes that are simultaneously excited at a given frequency. In regions of low modal overlap, individual resonances are easily resolved. The frequency that divides the two regions is similar to the Schroeder frequency [5] in room acoustics, where the response is dominated by isolated resonances below the transition frequency, and is best modeled statistically above the frequency where simultaneous modal excitation occurs. In the high-frequency region, the vibrations localize around the driving element [6]. The plate dimensions, stiffness and the amount of

<sup>1</sup>These lengths were used for the down-sampled inputs. The full-frequency spectrogram used  $[N = 512, 2048, 8192]$  to best align frequency resolution between the two versions.



**Figure 1:** 4-Channel multi-resolution spectrogram, downsampled by a factor of 30 and normalized using the provided normalization factor.



**Figure 2:** 4-Channel multi-resolution spectrogram, normalized using the provided normalization factor with no downsampling.

damping affect the amount of  $MO$ , with large, thin, compliant plates with heavy damping having the greatest  $MO$ .

For parameter identification, our model is configured to focus on the isolated resonances below the transition frequency, where individual modes are well-defined spatially and degeneracy is limited. The plate response above the transition frequency would be useful for identifying the input actuator location; however, since this is a known quantity for this problem, our model does not focus on high-frequency content. Given the ranges of the parameters defined by the problem, we computed that the recorded IR may be downsampled by a factor of 30 to a sample rate  $f_s = 1470$  Hz and still allow the observation of the first 20 resonances. At this lower sampling rate, the longest-length spectrogram allows for a frequency resolution  $f_{res}$  of 1.4355 Hz as given by (5), compared to a maximum resolution of 43.0664 Hz prior to downsampling.

$$f_{res} = \frac{f_s}{N} \quad (5)$$

### 3. RESULTS

Our approach produced several models that outperformed the baseline solution. We describe each model and its results here to provide additional insight to our process, but the submitted results include only our best-performing model: the downsampled spectrogram with a phase channel (M1-DsPh). Each model was trained on a desktop PC with an Intel i9-11900K processor with 64GB RAM and a NVIDIA GeForce RTX 3090 GPU. Each model was scheduled for 1000 epochs and a patience of 30 cycles. Adam optimization was used with a cosine-annealed learning rate with an initial value of  $1 \times 10^{-3}$ . Compute times are given in Table 1. The

**Table 1:** Model Performance and Computation Time, 200-IR Validation Set (5 sec)

| Model    | Param. Mean | NMSE | Epochs # | Training Time (min:sec) |
|----------|-------------|------|----------|-------------------------|
| Baseline | 0.0605      |      | N/A      | N/A                     |
| M1       | 0.0078      |      | 50       | 94:20                   |
| M1-Ds    | 0.0061      |      | 39       | 61:33                   |
| M1-DsPh  | 0.0058      |      | 42       | 63:54                   |

results discussed in this section were computed with the provided *eval.py* script on a dataset of 200 five-second random IRs.

The first developed model (M1) used the same architecture as discussed above, with the spectrograms computed using full-bandwidth IRs as shown in Fig. 2 and omitting the phase channel. This model beats the baseline by a large margin, but underperforms in comparison to both downsampled models when considering the number of training cycles and training time in addition to NMSE. Modal separation in this higher-frequency region is likely too poor to provide the model with any meaningful information about plate parameters. Thus, downsampling provides the model with the best resolution in the frequency range most likely to inform plate characteristics.

The second model (M1-Ds) is identical in architecture to M1 but operates on down-sampled input signals. This model performs notably better than the full-spectrum model and marginally worse than the downsampled model that includes phase information. As predicted, the model that includes phase information (M1-DsPh)

**Table 2:** Model performance on output location  $(x_o, y_o)$ .

| Model   | $x_o$ NMSE | $y_o$ NMSE |
|---------|------------|------------|
| M1-Ds   | 0.0035     | 0.0033     |
| M1-DsPh | 0.0033     | 0.0017     |

**Table 3:** Model Inference Time, 200-IR Validation Set (5-Sec)

| Model   | Mean Param<br>Inference Time (ms) | Real-Time Factor<br>( $\times 10^{-3}$ ) |
|---------|-----------------------------------|------------------------------------------|
| M1      | 6.48                              | 1.296                                    |
| M1-Ds   | 6.46                              | 1.292                                    |
| M1-DsPh | 6.47                              | 1.294                                    |

further improves the estimation of output location parameters  $(x_o, y_o)$  as shown in Table 2 compared to the models that omit phase data.

All models have very similar inference times for our 200 IR sample dataset, as shown in Table 3, so the increased accuracy of model M1-DsPh is achieved without a notable cost in inference time. Inference time is computed by measuring the elapsed time across the single PyTorch function call of our model, as shown in the pseudo code below.

```
start = time.time()
raw_output = model(spec_tensor
                    ).squeeze(0).cpu()
elapsed = (time.time() - start) * 1000.0
```

The computer used for inference was an Apple MacBook Air with an M2 processor and 16GB of RAM. We include the inference times for the reference data set in Table 4 as a control. However, it should be noted that this is a very small dataset size and may be influenced by background CPU activities during inference. We consider the 200 IR inference time a more accurate representation of system performance.

#### 4. CONCLUSION

The results of this work demonstrate the viability of leveraging machine learning models developed for image processing to solve audio system identification problems. High accuracy in physical parameter estimation may be achieved by converting audio signals to spectral images that are used as input features for the machine learning model. Input features to the ResNet-18 model may be further refined by employing phase information and restricting the frequency bandwidth to regions of low modal overlap. We hope this framework can be extended to empirical data for system identification on real plate reverb units.

#### 5. REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [2] C. Fuller, S. Elliott, and P. Nelson, Eds., *Active Control of Vibration*. London: Academic Press, 1996.

**Table 4:** Model Inference Time, 16-IR Official Dataset

| Model   | Mean Param<br>Inference Time (ms) | Real-Time Factor<br>( $\times 10^{-3}$ ) |
|---------|-----------------------------------|------------------------------------------|
| M1      | 44.24                             | 8.848                                    |
| M1-Ds   | 37.63                             | 7.526                                    |
| M1-DsPh | 14.56                             | 2.912                                    |

- [3] F. J. Fahy, *Sound and structural vibration: radiation, transmission and response*. Elsevier, 2007.
- [4] G. Rabbiolo, R. Bernhard, and F. Milner, “Definition of a high-frequency threshold for plates and acoustical spaces,” *Journal of Sound and Vibration*, vol. 277, no. 4-5, pp. 647–667, 2004.
- [5] M. R. Schroeder and K. Kuttruff, “On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima,” *The Journal of the Acoustical Society of America*, vol. 34, no. 1, pp. 76–80, 1962.
- [6] D. A. Anderson, M. C. Heilemann, and M. F. Bocko, “Measures of vibrational localization on point-driven flat-panel loudspeakers,” in *Proceedings of meetings on acoustics*, vol. 26, no. 1. Acoustical Society of America, 2016, p. 065003.