PEAK-RESIDUAL MODAL ESTIMATION WITH LEARNED CALIBRATION AND HIGH-BAND DENSITY CORRECTION

Doohyun Jung

Independent Researcher
jdoo0825@gmail.com

ABSTRACT

This paper describes two related submissions to Task B of the 1st DAFx Parameter Estimation Challenge. Both estimate modal frequency, decay, and gain directly from an unnormalised plate impulse response without using plate parameters, the excluded analytical modal-frequency law, or official-test ground truth. The primary system constructs a large candidate pool through prominence-graded spectral peak picking, iterative residual analysis, multi-view consensus, band-wise budgeting, and a learned file-level mode-count target. Raw decay and gain estimates are then corrected by a small mode-wise neural network that is not allowed to move frequencies or change the number of rows. A secondary variant addresses suspected high-frequency under-counting with a separately gated, non-oracle density-fill stage in the 6–10 kHz band. The paper reports development diagnostics, reproducibility information, and descriptive statistics for the 16 official outputs. The two variants expose a deliberate precision–recall trade-off: one preserves a visible spectral justification for every row, while the other tests bounded hidden-multiplicity augmentation in densely overlapped regions.

1. INTRODUCTION

Modal parameter estimation from an impulse response is straightforward when resonances are isolated, but becomes substantially harder when thousands of damped modes overlap. A single spectral maximum may represent one dominant mode, several neighbouring modes, or the combined shoulder of a dense modal cluster. The task is therefore not only to estimate the location and width of visible peaks, but also to infer how many physically meaningful modes are hidden beneath the measured response.

Task B of the DAFx Parameter Estimation Challenge asks participants to return, for each unnormalised impulse response, a table of modal frequencies, decay constants, and gains [1]. Classical approaches to exponentially damped sinusoids include sinusoidal modelling [2], linear-prediction and pole-estimation methods [3], and subspace or matrix-pencil methods [4]. These methods motivate joint parameter estimation, but their direct application at the modal densities and frequency span of the challenge can be computationally demanding and sensitive to model order.

The submitted approach instead separates three uncertainties. Frequency support is determined by a deterministic spectral front end. Local scale bias in decay and gain is handled by a constrained neural calibrator. Possible under-counting in the high band is treated as an optional post-processing hypothesis rather than being hidden inside the primary estimator. This produces two comparable outputs: a conservative peak-supported table and an augmented table that changes only high-band density.

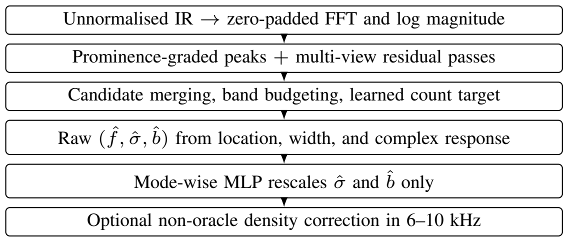

Figure 1: Combined processing pipeline. The final block is used only for the secondary submission.

2. TASK DEFINITION AND CONSTRAINTS

The measured discrete-time impulse response is modelled as a sum of damped sinusoids,

$$h[n] = \sum_{m=1}^{M} b_m e^{-\sigma_m n/f_s} \sin(2\pi f_m n/f_s), \tag{1}$$

where f_m is modal frequency, $\sigma_m > 0$ is the decay constant, b_m is modal gain, and f_s is the sample rate. For each input NPZ file, the estimator writes a frequency-sorted CSV file with columns `f0_ident`, `sigma_ident`, and `gain_ident`.

Only the unnormalised NPZ response is used at inference. Peak-normalised WAV files are not used because normalisation removes the absolute scale required for gain estimation. The code does not read plate parameters or official target modal tables, and it does not evaluate the excluded analytical modal-frequency or frequency-dependent decay relations. Learned components are fixed before official inference and are trained only on development responses generated with the public simulator.

3. PRIMARY PEAK-RESIDUAL ESTIMATOR

Figure 1 summarises the complete system. The primary submission stops after mode-wise calibration; the secondary submission includes the final density-correction block.

3.1. Candidate Discovery

A factor-eight zero-padded real FFT is computed, and candidate discovery operates mainly on the log-magnitude spectrum. Visible peaks are collected using a descending sequence of prominence thresholds. Early high-prominence passes capture isolated modes, while later low-prominence passes admit weaker structures. Minimum candidate spacing is frequency dependent: low-frequency candidates are kept farther apart and high-frequency regions are allowed to be denser.

A median spectral envelope is subtracted to form a residual curve. Accepted candidates are converted into soft Lorentzian-like masks, and the attenuated residual is searched again. Four residual passes are used, with subtraction strength multiplied by 0.75 after each pass. In addition to the base residual, deterministic smooth, sharp, sensitive, and strict views modify smoothing, gating, and attenuation. Nearby detections are merged, and the number of views supporting a candidate contributes to its score. Residual attenuation is intentionally heuristic: it is used to reveal shoulders and secondary ridges, not as an exact physical subtraction of already estimated modes.

3.2. Band Budgeting and Count Prediction

The interval 50–10,000 Hz is divided at 200, 500, 1000, 2000, 3500, 5000, and 7500 Hz. Each band receives a score that combines residual energy, residual-peak density, and log-magnitude variance. A file-level Ridge regressor predicts the desired number of modes from 37 signal-derived quantities, including strong and sensitive peak counts, spectral entropy and flatness, residual-energy ratios, candidate-pool size, score-tail statistics, and band-wise energy and density. Standardised, squared, transformed, and selected interaction terms yield 91 regression inputs.

For the official run, the predicted count is multiplied by a safety factor of 1.15, clipped to 700–7000 rows, rounded to the nearest 25, and limited by the available candidate pool. Candidates are then selected using the band budgets and global target together. The count model is not used to invent frequencies: it only determines how many signal-derived candidates survive.

3.3. Raw Frequency, Decay, and Gain

Each selected peak is parabolically refined in frequency. If f_L and f_R denote the local -3.01 dB crossings, the initial decay estimate is

$$\hat{\sigma}_{\mathrm{raw}} \approx \pi(f_R - f_L). \tag{2}$$

When valid crossings are unavailable, the estimator uses a nearby-frequency median of valid decay estimates. The complex spectrum is interpolated at the refined frequency to obtain $\hat{H}(f_m)$. Following the modal response convention used by the challenge baseline, gain is estimated as

$$\hat{b}_{\mathrm{raw}} = -2\ \mathrm{Im}\{\hat{H}(f_m)\}\frac{\hat{\sigma}_{\mathrm{raw}}}{f_s}\sin\left(\frac{2\pi f_m}{f_s}\right). \tag{3}$$

Thus the frequency placement, initial decay, initial gain, and row count remain tied directly to the measured response.

3.4. Mode-Wise Learned Calibration

Peak width and complex magnitude are systematically biased when modes overlap. A small multilayer perceptron (MLP) predicts multiplicative corrections rather than absolute parameters. Its checkpoint has 11 input channels: log frequency, log raw decay, log absolute raw gain, local mode density within ± 100 Hz, measured magnitude at the candidate, a five-band one-hot code with edges 50, 300, 1000, 3000, 6000, and 10,000 Hz, and a reserved prominence channel. Because the submitted v7.5 tables contain only the three required output columns, the prominence channel is identically zero in the submitted system. The network contains two 64-unit ReLU layers, dropout 0.3, and two outputs, for a total of 5058 trainable parameters.

Development candidates are matched to simulator ground truth by frequency. The targets are log ratios,

$$\Delta_\sigma = \log\frac{\sigma_{\mathrm{gt}}}{|\hat{\sigma}_{\mathrm{raw}}|}, \qquad \Delta_b = \log\frac{|b_{\mathrm{gt}}|}{|\hat{b}_{\mathrm{raw}}|}. \tag{4}$$

At inference, both outputs are clipped to $[-2, 2]$ and applied as

$$\hat{\sigma} = \hat{\sigma}_{\mathrm{raw}} e^{\Delta_\sigma}, \qquad \hat{b} = \mathrm{sign}(\hat{b}_{\mathrm{raw}})|\hat{b}_{\mathrm{raw}}|e^{\Delta_b}. \tag{5}$$

Frequency strings are copied unchanged, and the calibrator cannot add, delete, or reorder rows. This constraint keeps learning in the role of local bias correction.

4. HIGH-BAND DENSITY CORRECTION

The secondary submission addresses a specific failure hypothesis: in dense high-frequency regions, several rapidly decaying physical modes may collapse into fewer visible spectral structures. Its input is the calibrated modal table from the primary submission. A separate Ridge model predicts the target count in 6–10 kHz from six table-derived features: counts below 3 kHz, from 3–6 kHz, and from 6–10 kHz; total count; high-band fraction; and low-band fraction.

Let $\hat{M}_{\mathrm{hi}}$ be the predicted high-band target and M_{hi} the current high-band count. The requested number of additions is

$$N_{\mathrm{add}} = \mathrm{round}\Big[s\max\Big(0, \min(\hat{M}_{\mathrm{hi}}, cM_{\mathrm{hi}}) - M_{\mathrm{hi}}\Big)\Big], \tag{6}$$

with shrinkage $s = 1$ and maximum target ratio $c = 4$. Requests below 50 additions are ignored. Otherwise, existing high-band rows are sampled with replacement; copied frequencies are jittered uniformly within ± 20 Hz and clipped to the band. Only inserted rows are transformed:

$$\sigma_{\mathrm{new}} = 3\sigma_{\mathrm{sampled}}, \qquad b_{\mathrm{new}} = 2b_{\mathrm{sampled}}. \tag{7}$$

Original rows receive no intentional parameter transformation; CSV parsing and serialisation introduce only roundoff at machine precision. The method should therefore be interpreted as a bounded density correction, not as exact reconstruction of individual hidden modes.

5. EVALUATION AND OUTPUT ANALYSIS

5.1. Development Diagnostics

The global mode-count model was trained on 300 synthetic simulator responses, using 240 for training and 60 for holdout evaluation. Table 1 reports the stored diagnostics. The holdout set spans 928–14,092 modes. Its mean predicted count is close to the mean target, although the per-file absolute error remains non-negligible. The MLP was trained for at most 50 epochs with a file-level validation split. Its best validation mean-squared error (MSE) on the two log-ratio targets was 1.659 at epoch 47; the corresponding training MSE was 1.946. These values are training diagnostics, not the organisers' final Task B score.

Table 1: Development diagnostics for the learned components.

Quantity	Value
Count-model files: total / train / holdout	300 / 240 / 60
Holdout target count, min–mean–max	928–4551.8–14092
Holdout prediction, min–mean–max	944.8–4557.2–13416.7
Holdout mean / median absolute error	459.8 / 323.7
Holdout maximum absolute error	1339.5
MLP parameters	5058
Best epoch: training / validation MSE	47: 1.946 / 1.659

Table 2: Official-set output statistics over 16 impulse responses.

Output	Min	Median	Max	Total
Primary rows	3073	3365.5	3810	54785
Augmented rows	3073	4617.0	5374	71506
Added rows	0	1096.5	1773	16721
Primary 6–10 kHz	352	538.0	674	8329
Augmented 6–10 kHz	512	1521.0	2364	25050

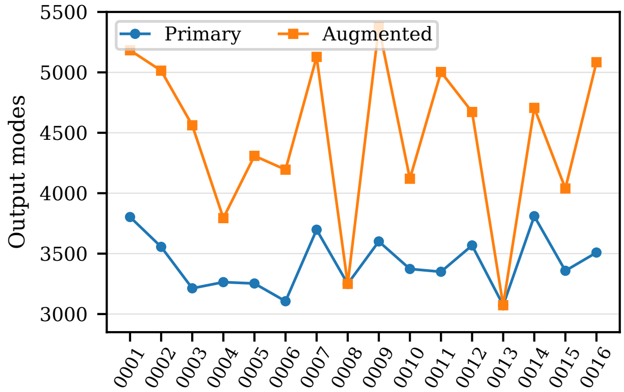

Figure 2: Per-file output counts. The augmented proposal differs from the primary proposal only through the bounded high-band filling stage.

5.2. Official Output Characteristics

The official set contains 16 unlabelled impulse responses. Both variants were produced in one deterministic CPU run under Windows 11 and Python 3.12. The primary pipeline required 5974.07 s in total, or 373.4 s per file on average. The complete secondary run required 5974.80 s; the density stage itself added approximately 0.73 s over all 16 files.

Table 2 and Fig. 2 describe the returned tables. The primary submission contains 54,785 rows, with a median of 3365.5 per response. The augmented submission contains 71,506 rows, an increase of 30.5%. All 16,721 inserted rows lie in the designated high band. Two responses receive no additions, while the largest increase is 1773 rows. The total number of 6–10 kHz rows rises from 8329 to 25,050.

Official target modal tables were not available to the estimator or included in the submission materials. Consequently, the statistics above document implementation behaviour and reproducibility rather than superiority under the organisers' final metric. The two variants represent different operating points. The primary output favours direct spectral support and lower false-positive risk. The augmented output favours recall in the most densely overlapped band, but accepts the possibility of over-filling and correlated inserted rows.

6. DISCUSSION

The main design choice is the separation of candidate discovery, parameter calibration, and count augmentation. The front end provides interpretable frequency evidence and makes the official inference deterministic. The MLP is deliberately unable to move a peak or alter table size, reducing the risk that development-data fitting creates an unconstrained modal distribution. The density stage is isolated as an optional hypothesis, so any change in the official score can be attributed primarily to high-band count behaviour rather than to a completely different estimator.

The principal limitation is that both variants remain peak centred. Residual masking can expose shoulders but cannot reliably separate several modes that form one smooth structure. The -3 dB width may also reflect local overlap rather than the decay of one physical mode, which is why calibration is necessary. Development matching introduces another limitation: the MLP is trained only on extracted candidates that can be paired with simulated modes. In the augmented variant, resampling preserves the local distribution of the current table, but the inserted frequencies are not independently supported by visible peaks. The fixed decay and gain multipliers are empirical development choices and may not generalise to all plate distributions.

The stored count checkpoint contains seven pairs of band energy and density, whereas the official extraction command computed eight allocation bands after adding a 7.5-kHz split. Consequently, the separate eighth pair was not consumed, and the last consumed pair had different bounds from those used during model fitting. This version mismatch is a limitation of the submitted run. Exact regeneration of the secondary proposal also requires the development modal tables used to fit its Ridge count model; those fitted coefficients are not contained in a separate checkpoint.

A stronger system would estimate count and parameters jointly while preserving numerical safeguards. Possible extensions include local matrix-pencil or subspace estimation in warped frequency bands [5], then reconstruction-based pruning; uncertainty-aware count prediction; and a learned proposal distribution conditioned on the measured residual rather than cloned modal rows. Once organiser-held labels or an equivalent independent test set are available, the most informative ablation would compare peak-only extraction, residual expansion, learned count control, MLP calibration, and density filling under the same official metric.

7. CONCLUSION

This paper presented a peak-residual modal estimator for DAFx Challenge Task B and a separately gated high-band density variant. The primary method combines deterministic multi-view candidate discovery, band-wise allocation, learned global count control, and constrained mode-wise decay/gain calibration. The secondary method adds a fast, non-oracle density correction in 6–10 kHz while preserving the primary rows up to machine-precision roundoff. The resulting pair makes the precision–recall trade-off explicit. Supplied code, primary-system checkpoints, and metadata document the submitted tables; exact regeneration of the secondary variant additionally requires the development tables used to fit its Ridge model.

8. REFERENCES

[1] M. Ducceschi and L. Gabrielli, "The 1st DAFx parameter estimation challenge: A benchmark for plate reverb system identification," in *Proceedings of the 29th International Conference on Digital Audio Effects (DAFx-26)*, Cambridge, MA, USA, 2026.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[3] R. Kumaresan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 833–840, 1982.

[4] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 5, pp. 814–824, 1990.

[5] O. Das and J. S. Abel, "Modal estimation on a warped frequency axis for linear system modeling," *arXiv preprint arXiv:2202.11192*, 2022.