A DUAL-STREAM FRAMEWORK COMBINING AUDIO SPECTROGRAM TRANSFORMER AND DYNAMIC MODE DECOMPOSITION FOR PLATE MODAL PARAMETER ESTIMATION

Yun Zhang[1] *Liangming Chen*[2]*, Zhenyu Guo*[1]*, Wei Liu*[1]*, and Gongping Huang*[1]

[1]Electronic Information School, Wuhan University, Wuhan, Hubei 430072, China
[2]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
yun_zhang@whu.edu.cn, clm267432@gmail.com,
zhenyuguo@whu.edu.cn, wei_liu@whu.edu.cn, gongpinghuang@whu.edu.cn

ABSTRACT

Plate reverberation is characterized by a dense distribution of resonant modes, which makes the estimation of modal parameters from observed responses a challenging inverse problem. To address this problem, we propose a physics guided dual stream framework that integrates an Audio Spectrogram Transformer (AST) with Dynamic Mode Decomposition (DMD). The AST branch models the global temporal and spectral structure of the impulse response, whereas the DMD branch extracts local descriptors associated with modal dynamics. The resulting representations are fused and processed by convolutional prediction heads to jointly estimate mode presence and the corresponding modal parameters. Experiments on the official validation set of Task B in the DAFx Challenge show that the proposed method reduces the overall relative error from 1.976 for the official baseline to 0.867. These results demonstrate that integrating local dynamic information derived from physical modeling with global transformer based representations substantially improves plate modal parameter estimation.

1. INTRODUCTION

Plate reverberation arises from thin plate vibration, with its impulse response governed by a dense set of resonant modes whose frequencies, decay rates, and gains depend on physical and geometrical properties, such as the plate material, geometry, boundary conditions, tension, and observation position [1, 2, 3, 4]. While forward simulation is well established, estimating modal parameters from observed responses remains difficult. Task B of the 1st DAFx Parameter Estimation Challenge addresses this inverse problem by requiring thousands of resonant modes to be estimated from a one-second impulse response, with each mode characterized by its natural frequency, decay constant, and modal gain. The official baseline relies on spectral peak picking followed by parameter estimation around each detected peak. While effective for prominent and well-separated resonances, this approach becomes unreliable for dense responses containing numerous closely spaced and overlapping modes.

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62471340 and in part by China Scholarship Council (CSC) under Grant 202506270089.

Dynamic Mode Decomposition (DMD) offers a complementary route to extracting local modal dynamics from time-series data [5, 6, 7, 8]. Given a sequence of observations from a dynamical system, DMD approximates the underlying evolution by a low-rank linear operator whose eigenvalues determine the oscillation frequency and decay rate, while the corresponding coordinates in the eigenvector basis represent the amplitudes. This makes DMD highly suitable for plate responses in the time-frequency domain, where each resonant mode manifests locally as an exponentially decaying trajectory across successive spectral frames. However, such a physical modeling lacks the global receptive field needed to interpret the holistic structure of plate reverberations. Recent studies have explored physics-aware learning frameworks that combine physical modeling with data-driven representation learning, such as differentiable modal simulation and inverse rendering [9, 10]. In parallel, large scale audio models pretrained on AudioSet [11], including the transformer based AST [12] and the CNN based PANNs [13], have shown strong capability in capturing the spectral structure and temporal evolution of audio signals.

To address the challenge of densely distributed and closely spaced modes in plate modal parameter estimation, this paper proposes a physics-guided dual-stream neural framework. The main contributions of this work are summarized as follows:

- **Physics-Guided Dual-Stream Framework:** We propose a unified end-to-end framework that couples a global audio spectrogram transformer branch with a local dynamic mode decomposition branch for joint mode detection and parameter regression.
- **DMD for Plate Modal Inversion:** We introduce dynamic mode decomposition into the plate reverberation inverse problem. By extracting local modal decay and amplitude descriptors from the time-frequency domain response, DMD provides explicit physical dynamic priors that complement the global transformer features, helping resolve closely spaced modes and stabilize parameter regression.
- **Performance Improvement:** Evaluated on the official validation set, the proposed method reduces the overall relative error from 1.976 for the official baseline to 0.867, with relative errors of 0.338 and 0.511 for natural frequencies and decay constants, respectively.

2. PROPOSED METHOD

2.1. Problem Formulation and Evaluation Band

Given an impulse response, let the ground-truth modal set contain M modes, represented by $(f_{0,m}, \sigma_m, b_m)_{m=1}^{M}$, where $f_{0,m}$, σ_m,

and b_m denote the natural frequency, decay constant, and modal gain of the m-th mode, respectively. The objective is to estimate a variable-length set of $\tilde{M}$ modes, represented by $(\hat{f}_{0,i}, \hat{\sigma}_i, \hat{b}_i)_{i=1}^{\tilde{M}}$. Following the challenge definition, only modes within the frequency range $[f_{\min}, f_{\max}]$ are considered. In the experiments, the frequency bounds are set to

$$f_{\min} = 20\,\text{Hz}, \qquad f_{\max} = 10\,000\,\text{Hz}. \tag{1}$$

Modes outside this band are ignored during training rasterization and evaluation.

2.2. Preprocessing

The displacement impulse response is amplitude-normalized before AST feature extraction and log-magnitude computation to improve numerical stability, while the original unnormalized response is used for DMD feature extraction to preserve the physical meaning of the estimated modal parameters.

To construct the input for the AST stream, the scaled response is converted into a log-mel filterbank spanning 256 time frames. A scalar log-energy feature is then computed and concatenated to the pretrained AST class token embedding.

To establish a unified discrete representation for both model targets and input features, we define a high-resolution logarithmic frequency grid with $K = 10\,000$ bins over the evaluation band:

$$c_k = f_{\min}\left(\frac{f_{\max}}{f_{\min}}\right)^{\frac{k-1}{K-1}}, \qquad k = 1, \ldots, K. \tag{2}$$

To facilitate supervised learning, ground-truth modes are projected onto the discrete log-frequency grid by assigning each mode to its nearest frequency bin c_k. When multiple modes collide in the same bin, only the mode with the largest absolute gain is retained. The resulting grid representation consists of a binary activity label $a_k \in 0, 1$ and the associated modal parameters (σ_k, b_k) for active bins, while inactive bins are filled with zeros.

The decay constants and modal gains span several orders of magnitude across different plate configurations, making direct regression challenging. To reduce the dynamic range, we first apply logarithmic transformation: $y_k = \ln(\sigma_k)$ and $z_k = \log_{10}|b_k|$. Subsequently, we perform z-score normalization:

$$\sigma_k^{\text{norm}} = \frac{y_k - \mu_y}{s_y}, \qquad b_k^{\text{norm}} = \frac{z_k - \mu_z}{s_z}, \tag{3}$$

where μ_y and μ_z denote the global means, and s_y and s_z the corresponding standard deviations, estimated from all active bins in the training set.

2.3. AST-DMD Dual-Stream Architecture

The proposed architecture consists of two parallel branches: an AST branch and a DMD branch, whose representations are fused on the 10 000-bin frequency grid, as illustrated in Fig. 1. The AST branch employs an Audio Spectrogram Transformer [12], initialized with AudioSet pretraining [11], to extract a global embedding from the input spectrogram. In parallel, the DMD branch provides physics-informed descriptors obtained from the Short-Time Fourier Transform (STFT) domain impulse response. The DMD descriptors are projected onto the logarithmic frequency grid for fusion with the AST embeddings. The fused representation is processed by one Conv1D layer with ReLU activation and dropout,

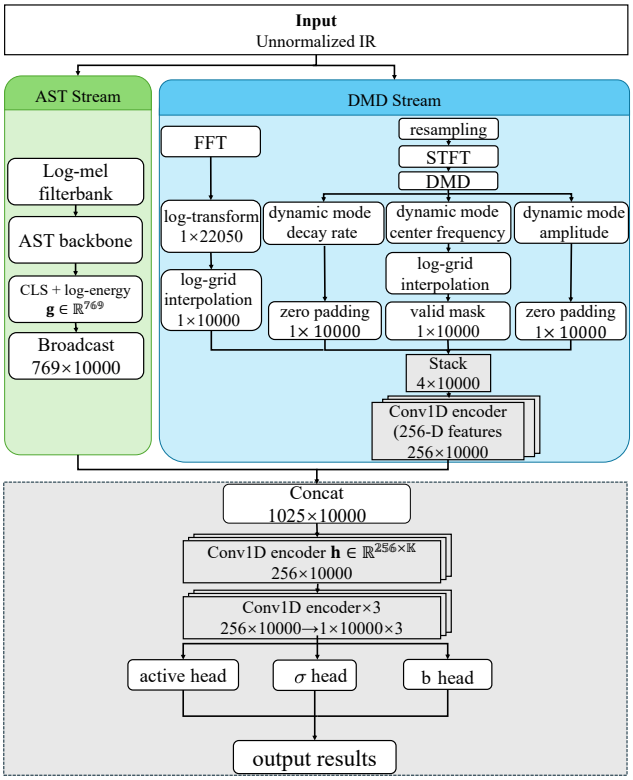

Figure 1: Overall architecture of the proposed framework.

producing the shared hidden feature map. Three parallel convolutional output heads are then applied on this shared feature map, generating predictions for mode activity, normalized decay, and normalized gain, respectively.

2.4. Local Mode Parameter Extraction via DMD

DMD is utilized to derive the localized dynamic-mode characteristics from the displacement impulse response. The DMD analysis is performed on the unnormalized response, since preserving the physical scale is important for estimating the dynamic-mode amplitudes. The response is first resampled to 48 kHz, followed by a transformation into the time-frequency domain using an STFT with a Kaiser window ($\beta = 1.9\pi$) configured with a frame size of 2048 samples and a hop length of 512 samples.

Let $\mathbf{Y} \in \mathbb{C}^{F \times L}$ denote the resulting complex STFT matrix, where F is the number of STFT frequency bins and L is the number of time frames. Instead of applying DMD to the full spectrogram, we perform local analysis along the frequency axis. For a selected STFT bin index m, a local spectrogram patch is extracted by collecting the 31 neighboring frequency bins from $m - 15$ to $m+15$. The first and last three time frames are discarded to reduce boundary effects. The resulting patch is denoted by $\mathbf{Y}_m \in \mathbb{C}^{31 \times T}$, where $T = L - 6$ is the number of retained time frames. The t-th column of $\mathbf{Y}_m$ is denoted by $\mathbf{y}_{m,t} \in \mathbb{C}^{31 \times 1}$ and represents the local spectral snapshot at time frame t.

For each local patch $\mathbf{Y}_m = [\mathbf{y}_{m,1}, \mathbf{y}_{m,2}, \ldots, \mathbf{y}_{m,T}]$, two time-shifted snapshot matrices are formed as $\mathbf{Y}_{m,1} = [\mathbf{y}_{m,1}, \ldots, \mathbf{y}_{m,T-1}]$ and $\mathbf{Y}_{m,2} = [\mathbf{y}_{m,2}, \ldots, \mathbf{y}_{m,T}]$. Exact DMD is then applied with a rank-R truncated SVD(Singular Value

Decomposition) where $R = 3$. This yields R DMD eigenvalues $\lambda_{m,r}$ and their corresponding complex DMD amplitudes $d_{m,r}$, with $r = 1, \ldots, R$ indexing the retained DMD components.

The local dynamic-mode decay rate and amplitude are computed as

$$\sigma_{\mathrm{DMD},m,r} = \ln(\lambda_{m,r}), \qquad g_{\mathrm{DMD},m,r} = |d_{m,r}|, \tag{4}$$

The center frequency of this dynamic mode is assigned as the center frequency f_m of the analyzed STFT neighborhood. Therefore, each retained DMD component produces a local dynamic-mode descriptor $(f_m, \sigma_{\mathrm{DMD},m,r}, g_{\mathrm{DMD},m,r})$. Note that $\sigma_{\mathrm{DMD},m,r}$ and $g_{\mathrm{DMD},m,r}$ are DMD-derived input features and are not identical to the target modal parameters.

The local DMD analysis is performed over selected STFT bins with a stride of 2 along the frequency axis. Since these descriptors are obtained on the linearly spaced STFT grid, they are projected onto the logarithmic frequency grid used by the network. Let $K = 10\,000$ denote the number of logarithmic frequency bins, and let c_k be the center frequency of the k-th log-frequency bin, where $k = 1, \ldots, K$. Each descriptor $(f_m, \sigma_{\mathrm{DMD},m,r}, g_{\mathrm{DMD},m,r})$ is assigned to the nearest log-frequency bin according to f_m. If multiple descriptors are mapped to the same log-frequency bin, only the one with the largest amplitude $g_{\mathrm{DMD},m,r}$ is retained. The mask is set to one for bins containing a mapped DMD descriptor and zero otherwise. Empty bins in the DMD decay-rate and amplitude channels are padded with zeros.

Note that all DMD-stream input features are precomputed offline to reduce training runtime overhead.

2.5. Dual-Stream Fusion

After both branches are aligned onto the shared logarithmic frequency grid, we encode the DMD features and fuse them with the AST outputs. The 4-channel raw DMD-stream tensor $\mathbf{S}_{\mathrm{DMD}} \in \mathbb{R}^{4\times K}$ is first processed by a Conv1D encoder. Then this encoder projects the low-dimensional physical features into a 256-channel latent space. The encoded DMD features are then concatenated with the aligned AST feature tensor $\mathbf{S}_{\mathrm{AST}} \in \mathbb{R}^{769\times K}$, yielding the fused feature tensor $\mathbf{S}_{\mathrm{fuse}} \in \mathbb{R}^{1025\times K}$. The fused feature tensor is subsequently refined by a Conv1D layer, producing a compact hidden feature map $\mathbf{h} \in \mathbb{R}^{256\times K}$. Finally, three parallel 1×1 convolutional output heads are applied on this map, and generate predictions for mode activity, normalized decay, and normalized gain, respectively.

During inference, the activity head predicts the probability of mode presence at each frequency bin, i.e., $\hat{p}_k$. A threshold τ is then applied to determine the active bins:

$$\hat{a}_k = \begin{cases} 1, & \hat{p}_k \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

For each active bin, the resonant frequency is assigned as the corresponding grid center frequency c_k. The modal decay rate and gain are subsequently recovered by inverting the logarithmic compression and z-score normalization defined in (3).

2.6. Loss Function

The proposed network comprises two parallel output branches. The classification branch performs binary activity detection at each frequency bin, while the regression branches estimate the corresponding normalized decay and gain parameters exclusively for the active bins. The loss function is accordingly decomposed into a focal loss for the classification task and a weighted Huber loss for the regression tasks, with the latter incorporating a frequency-dependent weighting scheme to reflect the physical dominance of low-frequency modes. The total training objective is the sum of these two components.

For the classification branch, let $a_k \in \{0, 1\}$ denote the ground-truth activity mask at bin k. The Focal Loss is formulated as

$$\begin{aligned} \mathcal{L}_{\mathrm{focal}} = -\frac{1}{K}\sum_{k=1}^{K}\Big[&\alpha(1-\hat{p}_k)^{\gamma}\log(\hat{p}_k)a_k \\ &+ (1-\alpha)\hat{p}_k^{\gamma}\log(1-\hat{p}_k)(1-a_k)\Big], \end{aligned} \tag{6}$$

where the focusing parameter and class balancing factor are set to $\gamma = 2$ and $\alpha = 0.35$, respectively.

For the regression branches, the loss is computed exclusively over the valid active bins ($a_k = 1$). To handle potential outliers in the predicted decay $\hat{\sigma}_k$ and gain $\hat{b}_k$, we employ the Huber loss with a transition threshold $\delta = 1$:

$$\mathcal{L}_{\mathrm{huber}}(x, y) = \begin{cases} \frac{1}{2}(x-y)^2, & |x-y| \leq \delta, \\ \delta\big(|x-y| - \frac{1}{2}\delta\big), & \text{otherwise.} \end{cases} \tag{7}$$

Since low-frequency modes typically contribute more strongly to the overall reverberant decay, we apply a normalized frequency-dependent weight v_k to the regression losses:

$$v_k = \frac{\exp(-c_k/500)}{\frac{1}{K}\sum_{j=1}^{K}\exp(-c_j/500)}, \tag{8}$$

where c_k is the pre-computed center frequency of bin k in Hz.

By integrating this dynamic weight, the regression objective is averaged over the active bins:

$$\begin{aligned} \mathcal{L}_{\mathrm{reg}} = \frac{1}{N_{\mathrm{act}}}\sum_{k=1}^{K} a_k v_k \Big[&\mathcal{L}_{\mathrm{huber}}(\hat{\sigma}_k, \sigma_k^{\mathrm{norm}}) \\ &+ \mathcal{L}_{\mathrm{huber}}(\hat{b}_k, b_k^{\mathrm{norm}})\Big]. \end{aligned} \tag{9}$$

The overall training objective is optimized by minimizing the combined loss:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{focal}} + \mathcal{L}_{\mathrm{reg}}. \tag{10}$$

3. EXPERIMENT

3.1. Experimental Setup

We train on 10 000 synthetic one-second impulse responses and evaluate on a held-out validation set of 1000 responses generated using the official data-generation code. The proposed system uses a 10 000-bin log grid over the analysis band in (1) and precomputed DMD sets for the training and validation splits. The AST backbone is initialized from AudioSet pretraining; it is then fine-tuned jointly with the grid heads using AdamW [14] with weight decay 10^{-4}, backbone learning rate 5×10^{-6}, head learning rate 10^{-4}, and a 3-epoch warm-up followed by cosine decay. Training runs for 100 epochs with batch size 4; the model checkpoint

with the lowest validation loss is selected for final inference with activity threshold $\tau = 0.411$.

Experiments were conducted on a Linux server with two Intel Xeon Platinum 8180 processors and 1 TB of system memory. The server is equipped with eight NVIDIA GeForce RTX 3090 GPUs.

The official Task B spectral-peak picking method is adopted as the baseline, with standard settings of 6 dB peak prominence and 2 Hz minimum peak spacing.

Performance is evaluated using the official challenge implementation of the Hungarian matching metric. The per-mode relative errors for frequency, decay, and gain are defined for each matched mode m as

$$\mathrm{re}_{\Omega}^{(m)} = \min\left(1, \frac{|\Omega_m^{(\mathrm{est})} - \Omega_m^{(\mathrm{ref})}|}{\Omega_m^{(\mathrm{ref})}}\right), \tag{11a}$$

$$\mathrm{re}_{\sigma}^{(m)} = \min\left(1, \frac{|\sigma_m^{(\mathrm{est})} - \sigma_m^{(\mathrm{ref})}|}{\sigma_m^{(\mathrm{ref})}}\right), \tag{11b}$$

$$\mathrm{re}_{b}^{(m)} = \min\left(1, \frac{|b_m^{(\mathrm{est})} - b_m^{(\mathrm{ref})}|}{b_m^{(\mathrm{ref})}}\right), \tag{11c}$$

where the superscripts $^{(\mathrm{est})}$ and $^{(\mathrm{ref})}$ denote estimated and ground-truth values, respectively. Let M denote the total number of true modes in the evaluation band. The per-plate relative errors are then averaged over all true modes:

$$\mathrm{RE}_{\Omega} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{re}_{\Omega}^{(m)}, \tag{12a}$$

$$\mathrm{RE}_{\sigma} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{re}_{\sigma}^{(m)}, \tag{12b}$$

$$\mathrm{RE}_{b} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{re}_{b}^{(m)}. \tag{12c}$$

The combined error without mode-count penalty is

$$\mathrm{RE}_0 = \frac{1}{3}\big(\mathrm{RE}_{\Omega} + \mathrm{RE}_{\sigma} + \mathrm{RE}_{b}\big). \tag{13}$$

Let $\tilde{M}$ be the number of estimated modes; the absolute mode-count mismatch is defined as $\Delta M = |M - \tilde{M}|$. The overall relative error incorporates a penalty for incorrect mode count and is given by

$$\mathrm{RE} = \mathrm{RE}_0 + \min\left(1, \frac{\Delta M}{M}\right). \tag{14}$$

3.2. Experimental Result

We evaluate the proposed AST-DMD method and the baseline on the official validation set of 1000 samples. Table 1 reports the mean and standard deviation of each metric, together with the overall score. The official baseline yields mean values of $\mathrm{RE}_{\Omega} = 0.986$, $\mathrm{RE}_{\sigma} = 0.994$, and $\mathrm{RE}_{b} = 0.990$, resulting in an overall RE of 1.976. In comparison, the proposed AST-DMD method reduces RE_{Ω} and RE_{σ} to 0.338 and 0.511, respectively, and lowers the overall RE to 0.867, corresponding to a relative reduction of 56%. The gain error remains high (RE_{b} = 0.907) but is still lower than the official baseline (0.990).

Table 1: Comparison on a validation set generated using the official data-generation code ($N = 1000$, mean $\pm$ std).

Metric	Baseline	Proposed
RE_{Ω}	0.986 ± 0.007	0.338 ± 0.178
RE_{σ}	0.994 ± 0.003	0.511 ± 0.084
RE_{b}	0.990 ± 0.006	0.907 ± 0.028
RE_0	0.990 ± 0.005	0.586 ± 0.094
ΔM	4867 ± 3492	1806 ± 2572
RE	1.976 ± 0.012	0.867 ± 0.259

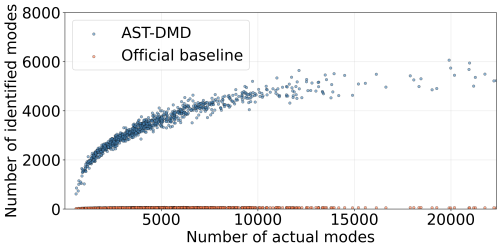

Figure 2: Comparison of the number of identified modes versus actual modes between the AST-DMD framework and the official baseline.

Figure 2 compares the estimated and ground truth numbers of modes for each of the 1,000 validation samples. The ground truth count includes all annotated modes within the frequency range from 20 Hz to 10 kHz. The orange markers represent the official baseline, which substantially underestimates the number of modes and typically detects fewer than 100 modes per sample. As a result, most baseline estimates are concentrated near the horizontal axis. In contrast, the blue markers representing the proposed method cover a much wider range and more closely track the variation in the ground truth counts. Quantitatively, the mean absolute mode count mismatch, denoted by ΔM, decreases from 4867 for the official baseline to 1806 for the proposed method. The corresponding standard deviations decrease from 3492 to 2572, indicating lower mode count errors and reduced variability across the validation set.

4. CONCLUSIONS

This paper presented a physics-guided dual-stream framework that integrates AST and DMD for plate modal parameter estimation. By combining global time frequency representations learned by a pretrained audio transformer with local modal dynamics extracted through DMD, the proposed method jointly exploits data driven representations and physically interpretable features. Experiments on the official validation set show that the proposed method substantially outperforms the official baseline, reducing the overall RE from 1.976 to 0.867. The proposed method also achieves more accurate estimates of natural frequencies and decay constants and substantially reduces the mismatch between the estimated and ground truth numbers of modes.

5. REFERENCES

[1] K. Arcas and A. Chaigne, "On the quality of plate reverberation," *Appl. Acoust.*, vol. 71, no. 2, pp. 147–156, 2010.

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

[3] J. O. Smith, "Music applications of digital waveguides," CCRMA, Department of Music, Stanford University, Tech. Rep. STAN-M-39, 1987.

[4] S. Bilbao, "Sound synthesis for nonlinear plates," in *Proc. Int. Conf. Digital Audio Effects (DAFx-05)*, 2005, pp. 243–248.

[5] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *J. Fluid Mech.*, vol. 656, pp. 5–28, 2010.

[6] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, "On dynamic mode decomposition: Theory and applications," *J. Comput. Dyn.*, vol. 1, no. 2, pp. 391–421, 2014.

[7] J. Kutz, S. Brunton, B. Brunton, and J. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*, ser. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2016.

[8] W. Liu, G. Huang, X. Liu, J. Jin, J. Chen, and J. Benesty, "Microphone array beamforming for speech enhancement based on dynamic mode decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2025, pp. 1–5.

[9] R. Diaz and M. Sandler, "Fast differentiable modal simulation of non-linear strings, membranes, and plates," in *Proc. Int. Conf. Digital Audio Effects (DAFx-25)*, 2025.

[10] X. Jin, C. Xu, R. Gao, J. Wu, G. Wang, and S. Li, "DiffSound: Differentiable modal sound rendering and inverse rendering for diverse inference tasks," in *Proc. ACM SIGGRAPH Conf. Papers*, 2024.

[11] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 776–780.

[12] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.