PROBING LOW-LEVEL ACOUSTIC ATTRIBUTE ENCODING IN CLAP AUDIO EMBEDDINGS

Héctor Martel, Joe Hennessy-Priest and Taemin Cho

BandLab Technologies, Singapore
hector.martel@bandlab.com

ABSTRACT

Audio foundation models are widely adopted as general-purpose feature extractors, yet the internal structure of their learned representations remains insufficiently understood. In this work, we analyze CLAP audio embeddings through a probing framework, studying the encoding of three fundamental perceptual dimensions: reverberation (RT60), loudness (LUFS), and spectral content, measured via spectral centroid (SC) and relative pitch (RP). Probes of increasing complexity are trained to predict each attribute from frozen embeddings across five datasets spanning noise, speech, monophonic musical notes, and music mixtures. Our primary finding is that all of these attributes are reliably recoverable from the CLAP embedding space across the examined datasets. Within this global picture, two encoding regimes emerge: RT60, LUFS, and RP are approximately linearly encoded, while SC requires nonlinear probes. Both regimes generalize across eight additional audio foundation models, with the notable exception that amplitude-invariant architectures discard loudness entirely by construction. The identified linear feature directions are geometrically consistent across datasets for RT60 and LUFS, while highly domain-specific for RP. Finally, we provide a qualitative demonstration of cross-modal consistency, showing that text embeddings of acoustic descriptors align geometrically with the identified RT60 feature direction.

1. INTRODUCTION

CLAP [1, 2] is a prominent audio-language foundation model that leverages contrastive learning to align audio and text in a shared embedding space. Trained on large collections of audio-text pairs, CLAP learns embeddings that capture both acoustic characteristics and semantic content, enabling applications such as audio retrieval, captioning [3, 4], audio quality assessment [5, 6], effect parameter estimation [7], room impulse response synthesis [8], and conditional audio generation [9, 10, 11]. The model has since been extended to incorporate temporal [12] and stereo [13] information.

Despite this widespread adoption, the internal representation structure of CLAP remains insufficiently understood. Probing studies of audio-language models have largely examined specific domains or model classes: early work probed speech embeddings for speaker and phonetic information [14], while more recent analyses of Large Audio-Language Models (LALMs), a model class that does not encompass CLAP, probe high-level perceptual attributes such as language, gender, and emotion of speech [15, 16].

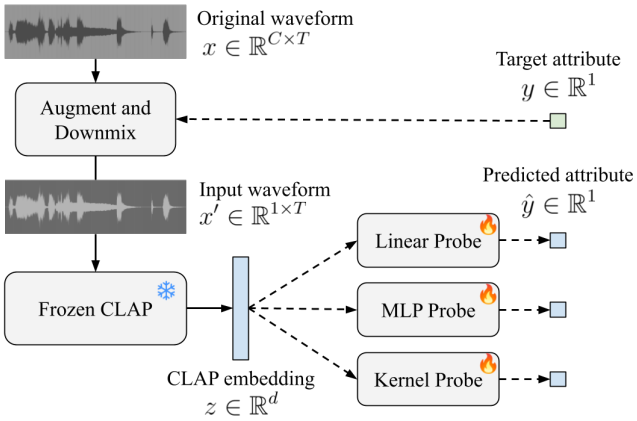


Figure 1: Overview of the probing methodology.

Interpretability approaches for acoustic models have favored qualitative concept alignment [17] over quantitative prediction of physical descriptors. Within the CLAP literature, studies have shown that high-level features such as timbre [18] and emotion [19] are effectively captured. However, low-level acoustic attributes have received far less attention. [20] shows that applying audio effects at varying intensities traces consistent trajectories in the CLAP embedding space. The authors exploit this property to improve embedding robustness. Conversely, FXEncoder++ [21] learns representations that capture audio effects independently of content. Text2FX [7] uses CLAP embeddings for audio effect parameter estimation via text prompts, implicitly assuming high cross-modal alignment and encoding of sufficient detail for fine-grained acoustic manipulation. In contrast, [6] proposes an Acoustic Context (ACX) Representation for audio quality enhancement and argues that CLAP embeddings fail to quantify noise or reverberation.

These mixed findings leave a central question open: which low-level acoustic attributes are encoded in CLAP embeddings, and how readily can they be recovered? Our work aims to address this gap through a systematic probing study.

2. METHODOLOGY

An overview of the probing methodology is presented in Figure 1. Each waveform $\mathbf{x} \in \mathbb{R}^{C\times T}$, with C channels and T samples, is transformed with an attribute-dependent data augmentation and downmixed to mono obtaining $\mathbf{x}' \in \mathbb{R}^{1\times T}$. $\mathbf{x}'$ is then encoded into $\mathbf{z} \in \mathbb{R}^d$ ($d = 512$), which serves as the sole input to all probes. The probes produce the predictions for single scalar attributes $\hat{y} \in \mathbb{R}^1$ and are trained on regression against the ground truth values $y \in \mathbb{R}^1$ used during the augmentation process. The

encoder is kept fully frozen: all learned parameters belong exclusively to the probe models. We use the audio encoder from LAION-CLAP [22, 23] checkpoint trained on speech, music, AudioSet, and LAION-Audio-630k[1], selected for its broad coverage across audio domains.

2.1. Target Acoustic Attributes

We investigate the encoding of low-level acoustic attributes in the CLAP embedding space, covering three fundamental perceptual dimensions: room acoustics, perceived loudness, and spectral content. All attributes are defined and computable across all datasets regardless of audio domain or musical instrumentation.

Reverberation Time (RT60). RT60 measures the time required for sound energy to decay by 60 dB following the cessation of a source, and is one of the most common descriptors of room acoustic conditions [24]. It is used here as a global measure of the degree of reverberation present in an audio sample, ranging from near-zero values in anechoic or close-miked recordings to several seconds in reverberant spaces.

Loudness (LUFS). Signal loudness is computed as integrated loudness following the ITU-R BS.1770 standard [25], using the `pyloudnorm` library[2]. Unlike peak amplitude or RMS energy, integrated LUFS accounts for frequency-dependent loudness perception, making it the de facto standard for loudness normalization in broadcast and music production.

Spectral Centroid (SC). The SC is the center of mass of the magnitude spectrum, $\mathrm{SC} = \sum_k f_k |X_k| / \sum_k |X_k|$, where k denotes the index of the frequency bin, f_k the corresponding frequency, and X_k its STFT magnitude [26]. We compute SC per frame from the STFT magnitude using a Hann window of length 2048 and hop size 512 samples, and then average across frames. We employ the SC rather than the fundamental frequency f_0 because it is well-defined for complex audio mixtures such as full songs, for which f_0 estimation is unreliable or ill-posed.

Relative Pitch (RP). The RP re-expresses SC on a logarithmic semitone scale as $\mathrm{RP} = 12 \log_2(\mathrm{SC}/f_{\mathrm{ref}})$, with $f_{\mathrm{ref}} = 440\,\mathrm{Hz}$ (A4), following the standard equal-tempered tuning relationship between frequency and pitch [27, 26]. RP and SC probe the same underlying spectral property from complementary scales, allowing us to assess whether encoding linearity depends on the scale used.

2.2. Probe Models

We evaluate three probe architectures of increasing complexity. All are intentionally lightweight so that performance differences reflect the embedding space rather than probe capacity.

Linear Probe. A single affine map $\hat{y} = \mathbf{w}^\top \mathbf{z} + b$, with $d{+}1 = 513$ parameters. The weight vector $\mathbf{w}$ defines the *feature axis*: the single direction in embedding space most predictive of the target. Strong performance indicates linear decodability, while poor performance indicates the need for non-linear encoding.

MLP Probe. A two-layer network $\hat{y} = \mathbf{W}_2\,\mathrm{GeLU}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) + b_2$ with 64 hidden dimensions (≈32.9k parameters). Shallow enough to avoid memorization, it captures non-linear relationships that the linear probe cannot. The MLP–Linear gap quantifies encoding non-linearity.

Kernel Probe. Kernel Ridge Regression (KRR) with an Radial Basis Function (RBF) kernel [28] $k(\mathbf{z}, \mathbf{z}') = \exp(-\gamma\|\mathbf{z} - \mathbf{z}'\|^2)$, $\gamma = 1.0$. Predictions are given by $\hat{y} = \sum_{j=1}^{M} \alpha_j k(\mathbf{z}, \mathbf{z}_j)$, where $\boldsymbol{\alpha} \in \mathbb{R}^M$ solves the $M \times M$ system $(\mathbf{K} + \lambda\mathbf{I})\boldsymbol{\alpha} = \mathbf{y}$, with $\mathbf{y} \in \mathbb{R}^M$ the target values for the M selected examples, and $\lambda = 10^{-3}$. M is capped at 10^4 randomly selected examples for tractability (KRR scales as $\mathcal{O}(M^3)$). Embeddings are L2-normalized before fitting. Since $\|\hat{\mathbf{z}} - \hat{\mathbf{z}}'\|^2 = 2(1 - \cos(\mathbf{z}, \mathbf{z}'))$ for unit vectors, the cosine RBF $k(\mathbf{z}, \mathbf{z}') = \exp\big(-2\gamma(1 - \cos(\mathbf{z}, \mathbf{z}'))\big)$ becomes the effective kernel. KRR makes no parametric assumptions and serves as a flexible non-linear reference without gradient-based optimisation sensitivity, providing an approximate ceiling for non-linear probe performance.

2.3. Evaluation Metrics

Probes are evaluated using three complementary metrics.

Mean Absolute Error (MAE) measures prediction error in the original unit of each feature, $\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$. It is directly interpretable but scale-dependent, preventing comparison across features with different units or dynamic ranges.

Coefficient of Determination (R^2) is the proportion of target variance explained by the model, $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$. $R^2 = 1$ is a perfect fit, $R^2 = 0$ matches a constant mean predictor, and negative values indicate worse-than-trivial performance. Unlike MAE, it is scale-free and accounts for task difficulty via the target variance.

Pearson Correlation (r) measures the linear association between predictions and ground truth, $r \in [-1, 1]$. Being insensitive to affine transformations, it is complementary to R^2: a high r with low R^2 indicates correct direction but miscalibrated scale. Conversely, a negative r indicates an encoding failure, as the probe recovers the attribute in the wrong direction on held-out data.

3. EXPERIMENTAL SETTINGS

3.1. Datasets

To test generalization across domains, we evaluate probes on five datasets spanning noise, speech, and music at increasing complexity, from monophonic notes to full music mixtures. All audio is resampled to 48 kHz to match the sample rate of LAION-CLAP.

White Noise (100k samples) is a synthetic control of generated white noise samples of 10 seconds each with constant amplitude envelope. Since all samples share the same flat spectral shape and carry no musical or semantic content, all variation in the embeddings is driven exclusively by the applied augmentation, making it the most controlled setting in our evaluation and serving as an upper bound on what a purely acoustic signal can encode. Labels are generated synthetically as described in Sec. 3.2.

NSynth [29] (306k samples) contains monophonic notes from 1,006 instruments at 16 kHz, spanning the full MIDI pitch range (21–108) at five velocity levels. Each sample is a 4-second clip in which a single note is held for three seconds and decays over the final second. It represents the simplest scenario in our evaluation, with a single note isolated from rhythmic or harmonic context.

VCTK-Corpus [30] (44k samples) is a speech dataset containing recordings from 110 English speakers with diverse accents. The recordings are clean, studio-quality speech with minimal background noise, making them well-suited for isolating acoustic attributes in our study.

MusDB18HQ [31] (34k samples) is a music source separation dataset containing 150 songs at 44.1 kHz. We use the full additive

[1] `laion-clap`: https://github.com/LAION-AI/CLAP

[2] `pyloudnorm`: https://github.com/csteinmetz1/pyloudnorm

mix, segmented into 10-second chunks with 1-second overlap. To avoid data leakage, the training, validation, and test sets are split by song. MusDB18HQ represents a more complex musical scenario than NSynth, with multiple instruments, polyphony, and realistic production arrangements. The mixes are raw recordings without compression, equalization, or mastering, making low-level acoustic attributes directly observable.

SonicMaster [32] (525k samples) comprises songs across 10 balanced genre groups. The dataset contains ground truth mastered tracks and degraded versions with applied audio effects (e.g. equalization, dynamics). We consider only the ground truth mastered tracks and split them by song to ensure no data leakage. This dataset represents the most production-realistic and largest-scale scenario in our benchmark.

For all datasets, targets are measured as per Sec. 3.2 and we use 80%/10%/10% splits for training, validation, and testing. The splits of VCTK and NSynth are speaker-disjoint and instrument-disjoint, respectively.

3.2. Feature Generation

To ensure that each acoustic attribute spans a balanced distribution of values across its perceptually relevant range, the original datasets are augmented by applying a single targeted perturbation per sample. The embedding and attribute value are then computed from the perturbed signal. Each attribute is augmented independently (i.e., only one perturbation is applied per sample) so that attributes such as RT60 and LUFS cannot co-vary as a result of the augmentation procedure, and any correlation between their learned probe directions reflects embedding structure rather than a data pipeline confound. Following this regime we generate a full copy of each dataset per attribute.

RT60: A target reverberation time is uniformly sampled as $t \sim \mathcal{U}(0.0,\ 2.0)$ s. A synthetic room impulse response (RIR) is generated using `gpuRIR`[3]. Room dimensions are sampled uniformly per axis in $[4, 12]$ m, and wall reflection coefficients are derived analytically from t via the Sabine equation [33]. The RIR is convolved with the dry waveform, and the output is RMS-normalized to the level of the input to prevent loudness from acting as a proxy cue for RT60.

LUFS: A target loudness level is uniformly sampled as $l \sim \mathcal{U}(-40,\ -10)$ LUFS. The signal is then normalized to l using the loudness normalization function provided by `pyloudnorm`, ensuring that the training distribution covers a wide and balanced range of perceived loudness levels.

SC: Two procedures are used depending on the dataset. For White Noise, a target SC value is sampled as $f_c \sim \mathcal{U}(500,\ 5000)$ Hz and the noise is filtered with a bandpass filter centered at f_c with a bandwidth drawn from $b \sim \mathcal{U}(0.25,\ 4)$ octaves, controlling the spectral center of mass. For all other datasets, a fractional pitch shift $p \sim \mathcal{U}(-6,\ 6)$ semitones is uniformly sampled and applied to the waveform via the `torch_audiomentations`[4] phase vocoder. This creates a smooth distribution of pitch in all datasets, breaking the quantised structures present in music datasets (e.g. NSynth contains only integer pitch values, which we found to hinder regression performance). The two strategies are not interchangeable by design: bandpass filtering is appropriate only for spectrally flat signals, while pitch shifting preserves content. Since probes are always trained and evaluated within the same dataset, no transfer across augmentation strategies is assumed or performed. In both cases, the SC is computed as defined in Sec. 2.1 using the `SpectralCentroid` transform from `torchaudio`[5], yielding a single scalar in Hz.

RP: No additional augmentation is applied for RP. The attribute value is obtained from the SC value following the method described in Sec. 2.1.

[3] `gpuRIR`: https://github.com/DavidDiazGuerra/gpuRIR

[4] `torch_audiomentations`: https://github.com/iver56/torch-audiomentations

3.3. Training Details

The Linear and MLP probes are trained for 100 epochs with a batch size of 256 to minimize the MSE objective with AdamW optimizer, using a learning rate of 1×10^{-3} and weight decay of 1×10^{-3}. Early stopping is applied based on the validation loss, with a patience of 10 epochs. The Kernel probe is fitted in a single training iteration. For the largest dataset (SonicMaster), results were verified to be stable when increasing M from 10^4 to 10^5, confirming the cap does not constrain performance. All reported results are averaged over 10 independent runs with different random seeds.

4. RESULTS

Results are shown in Table 1. A global trend can be observed: non-linear probes (MLP, Kernel) consistently outperform the Linear probe across all features and datasets, with the performance margin growing from modest gains for RT60 and LUFS, to substantial gains for SC, and moderate gains for RP. This implies that the geometry of CLAP's encoding varies considerably across features, which we analyze in more detail below.

RT60 is strongly and nearly linearly encoded. The Linear probe achieves $R^2 \geq 0.67$ and $r \geq 0.82$ across all datasets, and $R^2 \geq 0.92$ on White Noise and VCTK-Corpus. MAE ranges from 0.09 s on White Noise to 0.25 s on NSynth. Non-linear probes provide consistent improvements: gains are modest on White Noise and VCTK-Corpus ($\Delta R^2 \leq 0.04$) but more substantial on NSynth, where the Kernel probe reaches $R^2 = 0.81$ vs. the Linear probe's $R^2 = 0.67$. The Kernel probe reaches $R^2 \geq 0.75$ across all datasets ($R^2 = 0.99$ on White Noise), confirming a predominantly linear underlying structure.

LUFS is also strongly and approximately linearly encoded. The Linear probe achieves $R^2 \geq 0.76$ and $r \geq 0.88$ (MAE: 2.1–3.5 LUFS across real datasets). Non-linear probes bring moderate improvements ($R^2 \geq 0.88$, $r \geq 0.94$, MAE: 1.4–2.3 LUFS), with MLP and Kernel performing comparably across most datasets.

SC results substantially differ from RT60 and LUFS. The Linear probe performs at or below chance on four of five datasets. On VCTK-Corpus and MusDB18HQ (†) with $r = -0.29$ and $r = 0.04$ respectively, on SonicMaster $R^2 = -0.06$ with $r = 0.35$, and on White Noise (†) with $r = 0.51$, indicating the correct direction was partially identified but at a severely miscalibrated scale. NSynth is a notable exception: the Linear probe achieves $R^2 = 0.43$ ($r = 0.68$), a partial linear recovery not observed on any other dataset, consistent with its simplified acoustic content (isolated monophonic notes) producing a particularly clean SC signal in the embedding space. Non-linear probes recover SC substantially better across all datasets: the MLP reaches $R^2 = 0.84$ on SonicMaster, and the Kernel probe achieves the best results

[5] `torchaudio`: https://github.com/pytorch/audio

Table 1: Test-set probing results across datasets, averaged over 10 independent seeds. Bold indicates the best or joint-best result per (dataset, feature). †: probe failure ($R^2 < -1$); negative R^2 values above this threshold are reported explicitly. Standard deviations of R^2 across seeds are ≤ 0.01 for RT60 and LUFS, ≤ 0.05 for SC (MLP and Kernel only, as Linear probe fails on SC), and ≤ 0.04 for RP.

		White Noise			NSynth			VCTK-Corpus			Musdb18HQ			SonicMaster		
Feature	Probe	MAE ↓	R^2 ↑	r ↑	MAE ↓	R^2 ↑	r ↑	MAE ↓	R^2 ↑	r ↑	MAE ↓	R^2 ↑	r ↑	MAE ↓	R^2 ↑	r ↑
RT60	Linear	0.09	0.95	0.98	0.25	0.67	0.82	0.12	0.92	0.96	0.21	0.77	0.88	0.21	0.76	0.87
	MLP	0.07	0.97	0.99	0.21	0.75	0.87	0.10	0.94	0.97	**0.20**	**0.79**	**0.89**	**0.19**	**0.80**	**0.89**
	Kernel	**0.04**	**0.99**	**1.00**	**0.18**	**0.81**	**0.90**	**0.09**	**0.95**	**0.98**	0.21	0.75	0.88	0.19	0.79	**0.89**
LUFS	Linear	0.5	**0.99**	**1.00**	3.0	0.80	0.90	2.1	0.92	0.96	3.5	0.76	0.88	2.1	0.90	0.95
	MLP	0.3	**1.00**	**1.00**	**2.2**	**0.89**	**0.95**	**1.5**	**0.95**	**0.98**	2.1	0.90	0.95	**1.4**	**0.96**	**0.98**
	Kernel	**0.2**	**1.00**	**1.00**	2.3	0.88	0.94	**1.5**	**0.95**	**0.98**	**1.9**	**0.92**	**0.96**	1.5	0.95	**0.98**
SC	Linear	1451	†	0.51	183	0.43	0.68	1193	†	-0.29	2930	†	0.04	729	-0.06	0.35
	MLP	119	0.98	0.99	**79**	**0.90**	**0.95**	160	0.82	0.91	574	0.40	0.67	**278**	**0.84**	**0.92**
	Kernel	**34**	**1.00**	**1.00**	94	0.85	0.93	**129**	**0.89**	**0.95**	**467**	**0.60**	**0.79**	**279**	**0.84**	**0.92**
RP	Linear	1.44	0.97	0.99	4.31	0.73	0.86	2.34	0.75	0.88	3.83	0.36	0.64	2.12	0.82	0.91
	MLP	0.52	**1.00**	**1.00**	**2.63**	**0.90**	**0.95**	**1.66**	**0.87**	**0.94**	2.78	0.64	0.81	**1.55**	**0.90**	**0.95**
	Kernel	**0.21**	**1.00**	**1.00**	3.22	0.85	0.92	1.69	**0.87**	**0.94**	**2.63**	**0.70**	**0.85**	1.79	0.88	0.94

overall ($R^2 \geq 0.60$, $r \geq 0.79$, MAE ≤ 467 Hz). The predominant failure of the Linear probe and consistent success of non-linear probes evidences that SC generally lies on a curved manifold that a single linear projection cannot recover, with NSynth as a boundary case under simplified acoustic conditions.

RP is linearly encoded across most datasets, though with varying strength. The Linear probe achieves strong performance on White Noise ($R^2 = 0.97$), SonicMaster ($R^2 = 0.82$, $r = 0.91$), and VCTK-Corpus ($R^2 = 0.75$, $r = 0.88$), moderate performance on NSynth ($R^2 = 0.73$, $r = 0.86$), and substantially weaker performance on MusDB18HQ ($R^2 = 0.36$, $r = 0.64$). The weaker MusDB18HQ result is consistent with both its smaller training set size and the greater embedding variation introduced by full polyphonic mixes, which reduces the proportion of variance attributable to a single linear direction. In all cases, $r > 0$ confirms that the linear probe always identifies the correct direction in embedding space. This stands in sharp contrast to SC, a monotone transformation of RP yet one for which the Linear probe is essentially uninformative on most datasets: SC yields $r = -0.29$ on VCTK-Corpus and $r = 0.04$ on MusDB18HQ, with the negative value on VCTK-Corpus indicating that the Linear probe recovers SC in the wrong direction entirely. Non-linear probes recover RP reliably across all datasets ($R^2 \geq 0.64$, $r \geq 0.81$, with $R^2 = 1.00$ on White Noise), confirming that RP information is present regardless of poor linear decodability. Two factors may explain the SC/RP divergence: the log transform produces a more uniform target distribution that is intrinsically easier to fit linearly, and because CLAP operates on log-mel spectrograms, pitch shifts may be preserved as a near-linear direction in embedding space.

4.1. Feature Axis Analysis

All figures and numerical results correspond to the training run with the median test R^2 across the 10 independently seeded runs.

To quantify the geometric alignment of learned feature directions across datasets, we compute the pairwise cosine similarity $\cos(\mathbf{w}_i, \mathbf{w}_j) = \mathbf{w}_i^\top \mathbf{w}_j / (\|\mathbf{w}_i\| \cdot \|\mathbf{w}_j\|)$ between each pair of weight vectors trained independently on datasets i, j. The corresponding angle is $\theta = \arccos(|\cos(\mathbf{w}_i, \mathbf{w}_j)|) \in [0°, 90°]$. The absolute value accounts for the sign indeterminacy of the linear probe weight vector: $\mathbf{w}$ and $-\mathbf{w}$ define the same feature axis. For reference, in $d = 512$ dimensions, two random unit vectors have expected absolute cosine similarity $\mathbb{E}[|\cos|] \approx \sqrt{2/(\pi d)} \approx 0.035$ ($\theta \approx 88°$) for large d, i.e., nearly orthogonal by chance.

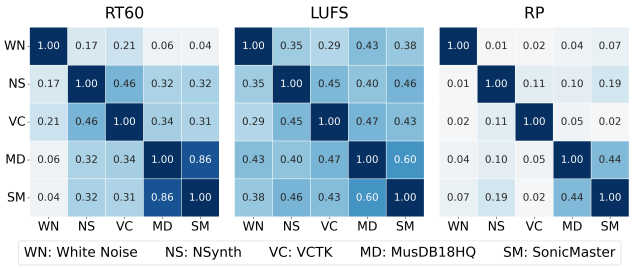
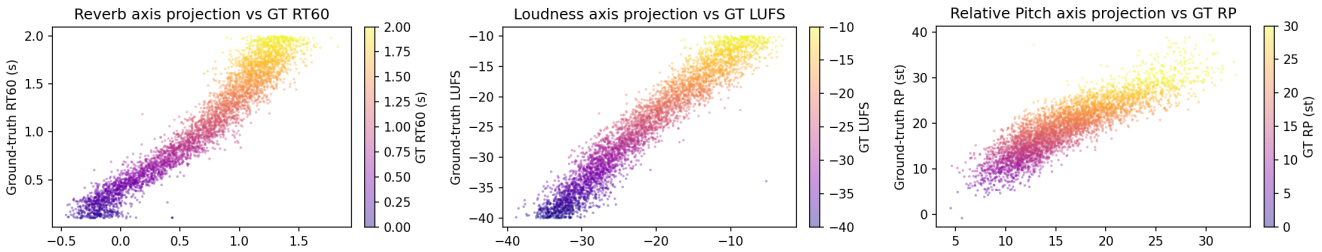


Figure 2: Pairwise cosine similarity of linear probe weight vectors $\mathbf{w} \in \mathbb{R}^{512}$ trained independently on each dataset, for RT60 (left), LUFS (center), and RP (right).

As shown in Figure 2, RT60 and LUFS feature axes are consistent across datasets. The RP axis, in contrast, is highly domain-dependent.

RT60 axes show moderate cross-domain alignment ($\cos \in [0.17, 0.46], \theta \in [62°, 80°]$). MusDB18HQ–SonicMaster is the strongest pair ($\cos = 0.86$, $\theta \approx 31°$), attributable to their shared music-mixture domain. NSynth and VCTK-Corpus align more strongly with each other ($\cos = 0.46, \theta \approx 62°$) than either does with the complex music mixtures ($\cos \in [0.31, 0.34]$, and $\theta \in [70°, 72°]$). Both are single-source datasets despite belonging to different domains, suggesting that content complexity is a strong organizing principle. Pairs involving White Noise fall at the lower end ($\cos \in [0.04, 0.21], \theta \in [78°, 88°]$), producing a probe direction with no content-driven components. This is consistent with the absence of salient content in the White Noise dataset. Excluding White Noise, all pairs substantially exceed the chance baseline. This is consistent with, though not definitive evidence of, a common RT60 direction across domains.



(a) RT60 axis projection vs. ground truth. **(b)** LUFS axis projection vs. ground truth. **(c)** RP axis projection vs. ground truth.

Figure 3: VCTK test-set embeddings projected onto the linear-probe weight vector (x-axis) vs. ground-truth value (GT, y-axis), colored by target value. Both RT60 and LUFS show a monotone band with mild curvature, confirming a predominantly linear structure consistent with $R^2 = 0.92$ for the Linear probe on VCTK-Corpus (Table 1). RP shows a more pronounced curve, consistent with a weaker linear encoding ($R^2 = 0.75$). SC is omitted as the Linear probe fails on VCTK-Corpus (†), making the projection uninformative.

LUFS axes are uniformly consistent ($\cos \in [0.29, 0.60]$, $\theta \in [53°, 73°]$), including pairs involving White Noise. The higher cross-domain consistency relative to RT60 is consistent with gain scaling being a globally uniform transformation that dominates the embedding displacement regardless of content. Residual variation nonetheless indicates that domain-specific components remain present in all estimated axes.

RP presents a markedly different picture: with the exception of MusDB18HQ–SonicMaster ($\cos = 0.44$, $\theta = 64°$), while most pairs are near or at the chance baseline, indicating near-orthogonal axes across domains. The one coherent pair shares a common music-mixture domain, suggesting that a common RP direction is recoverable only within a domain. Unlike RT60 and LUFS, whose global transformations leave a partially consistent trace across domains, the RP axis is dominated by domain-specific embedding structure and does not transfer across speech, isolated notes, noise, and music, even though RP is well-encoded and linearly recoverable within each domain individually.

Figure 3 illustrates what projecting onto a linear probe weight vector looks like in practice, using VCTK-Corpus as a representative dataset. Each embedding of the test set is projected onto the scalar $p = \mathbf{w}^\top \mathbf{z}/\|\mathbf{w}\|$ and plotted against its ground-truth value.

4.2. Effect of Training Set Size

To understand how probe performance scales with the amount of training data, we train Linear probes on progressively smaller subsets of SonicMaster and measure performance as a function of training set size. SonicMaster is chosen for this study as it is the largest and most domain-diverse dataset in our benchmark, making it a suitable basis for a data efficiency analysis. SC is excluded since the Linear probe fails at all training sizes. The Linear probe is used here by design: with only $d+1$ parameters, its saturation curve directly reflects the sample complexity of estimating a single direction in the embedding space, without confounds from probe capacity or gradient-based optimisation sensitivity that would arise with the MLP or Kernel probes.

Table 2 shows markedly different data efficiency profiles across features. RT60 is robust to limited data: the Linear probe achieves $R^2 = 0.73$ with as few as 4.2k training examples (ratio 0.01), and performance saturates at $R^2 = 0.76$ from ratio 0.05 onward. This suggests that the RT60 direction in the embedding space is high-variance and salient. LUFS and RP are considerably more data-hungry. LUFS improves monotonically from complete failure at ratio 0.01 to saturation at ratio 0.50 ($R^2 = 0.89$). RP remains negative at ratio 0.02 ($R^2 = -0.71$) before recovering from ratio 0.05 onward and saturating at ratio 0.50 ($R^2 = 0.79$). At least 21k examples (ratio 0.05) are needed before the probe first yields positive R^2, and saturation is reached at 210k examples (ratio 0.50).

Table 2: Linear probe test-set MAE / R^2 on SonicMaster at varying training-set sizes (val/test splits always use the full set). †: probe failure ($R^2 < -1$); negative R^2 values above this threshold are reported explicitly. SC is omitted as the Linear probe fails at all sizes. Pearson r is omitted from this and the following table as it is monotonically related to R^2 on a fixed test set.

		RT60		LUFS		RP	
Ratio	N ($\times 10^3$)	MAE	R^2	MAE	R^2	MAE	R^2
0.01	4.2	0.22	0.73	17.6	†	22.6	†
0.02	8.4	0.22	0.75	5.7	0.34	6.5	−0.71
0.05	21.0	**0.21**	**0.76**	3.7	0.72	3.4	0.51
0.10	42.1	**0.21**	**0.76**	2.7	0.85	2.6	0.71
0.20	84.1	**0.21**	**0.76**	**2.3**	0.88	**2.2**	0.78
0.50	210.0	**0.21**	**0.76**	**2.3**	**0.89**	**2.2**	**0.79**
1.00	421.0	**0.21**	**0.76**	**2.3**	**0.89**	**2.2**	**0.79**

4.3. Generalization to Other Models

To assess whether the encoding patterns observed in CLAP are specific to this model or arise more broadly from large-scale audio pretraining, we repeat the linear probe evaluation on the following additional audio embedding models: MS-CLAP versions of 2022 [1] and 2023 [2], MERT [34], Wav2Vec2 [35], WavLM (Base and Large) [36], Whisper [37], and VGGish [38]. As most of these models are trained and used in downstream tasks in the speech domain, the natural choice for this study is VCTK-Corpus.

Embeddings are extracted via `fadtk` [39][6] using the same training and evaluation protocol as Sec. 2.2, with probe dimensionality adapted to each model's embedding size d (MLP hidden size fixed at 64), and audio resampled to each model's sample rate.

Table 3 shows that the global trend from Sec. 4 holds across all models for RT60, SC, and RP. LUFS, however, varies substantially by architecture.

[6] `fadtk`: https://github.com/microsoft/fadtk

Table 3: Multi-embedder probe evaluation on VCTK-Corpus. MAE ↓, R^2 ↑; best or joint-best probe per (embedder, feature) in **bold**. †: probe failure ($R^2 < -1$); negative R^2 values above this threshold are reported explicitly.

		RT60		**LUFS**		**SC**		**RP**	
Embedder	Probe	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
MS-CLAP 2022 (1024)	Linear	0.14	0.89	4.6	0.54	233	0.60	**2.2**	**0.78**
	MLP	0.13	0.90	**4.2**	**0.59**	168	0.82	**2.2**	**0.78**
	Kernel	**0.12**	**0.92**	4.3	0.58	**167**	**0.82**	**2.2**	**0.78**
MS-CLAP 2023 (1024)	Linear	0.08	0.96	1.7	0.93	213	0.67	**1.5**	0.89
	MLP	**0.07**	**0.97**	**1.6**	**0.94**	121	0.90	**1.5**	**0.90**
	Kernel	**0.07**	**0.97**	1.7	0.93	**118**	**0.91**	**1.5**	0.89
MERT (768)	Linear	0.10	0.94	**7.4**	-0.02	494	-0.79	1.6	0.88
	MLP	0.10	0.95	**7.4**	**-0.01**	119	0.91	1.3	0.92
	Kernel	**0.08**	**0.96**	8.2	-0.29	**93**	**0.94**	**1.2**	**0.93**
Wav2Vec2 (768)	Linear	0.11	0.93	**7.5**	**0.00**	373	-0.03	2.2	0.79
	MLP	0.11	0.93	**7.5**	**0.00**	160	0.84	**1.7**	0.86
	Kernel	**0.09**	**0.95**	8.8	-0.50	**126**	**0.90**	**1.7**	**0.87**
WavLM-B (768)	Linear	0.12	0.91	3.8	0.68	496	-0.82	1.8	0.84
	MLP	0.11	0.93	**3.0**	**0.81**	136	0.88	**1.4**	**0.90**
	Kernel	**0.10**	**0.94**	3.1	0.79	**108**	**0.93**	**1.4**	**0.90**
WavLM-L (1024)	Linear	0.07	0.97	**7.5**	**-0.01**	490	-0.87	1.7	0.86
	MLP	0.07	0.97	**7.5**	**-0.01**	119	0.91	**1.4**	**0.91**
	Kernel	**0.05**	**0.98**	8.3	-0.31	**103**	**0.93**	**1.4**	**0.91**
Whisper (512)	Linear	0.10	0.94	1.1	0.97	426	-0.31	2.0	0.82
	MLP	0.10	0.95	0.8	0.98	154	0.85	**1.5**	**0.89**
	Kernel	**0.09**	**0.96**	**0.8**	**0.99**	**119**	**0.91**	**1.5**	**0.89**
VGGish (128)	Linear	0.15	0.87	4.0	0.68	1130	†	2.4	0.74
	MLP	0.11	0.93	2.6	0.86	168	0.83	**1.9**	**0.85**
	Kernel	**0.09**	**0.95**	**2.2**	**0.89**	**146**	**0.86**	**1.9**	**0.85**

RT60 is strongly and linearly encoded across all models (Linear probe $R^2 \geq 0.87$), which confirms that this trend generalizes beyond LAION-CLAP.

LUFS is well encoded in Whisper ($R^2 = 0.97$), which yields the best linear result across all models, and MS-CLAP 2023 ($R^2 = 0.93$). WavLM-Base and VGGish show weaker linear encoding ($R^2 = 0.68$), and non-linear probes improve performance ($R^2 = 0.81$ and $R^2 = 0.89$, respectively). MS-CLAP 2022 shows substantially weaker LUFS encoding ($R^2 = 0.54$ linear, 0.59 best probe) despite sharing the same contrastive audio-text objective as the 2023 version. Given that the two checkpoints differ in both backbone architecture (CNN14 vs. HTSAT-tiny) and training data scope, the precise cause of this difference cannot be isolated from the available evidence. All probes yield $R^2 \approx 0$ or below on MERT, Wav2Vec2 and WavLM-Large, indicating that loudness is absent from those representations rather than merely non-linearly encoded. In all cases, the failure is explained by architectural amplitude invariance: Wav2Vec2 and WavLM-Large normalize inputs to zero mean and unit variance before encoding, while MERT applies it in its convolutional feature extractor, which removes global amplitude scaling from all downstream activations.

SC remains non-linearly encoded across all non MS-CLAP models ($R^2 \leq 0$). Both MS-CLAP versions achieve partial linear recovery ($R^2 = 0.67$ and 0.60 for the 2023 and 2022 checkpoints respectively). This result is remarkable: the checkpoints differ in both audio encoder (CNN14 vs. HTSAT-tiny) and training data, yet exhibit partial linear SC recovery that no other model replicates.

RP is strongly and linearly encoded across all models ($R^2 \geq 0.74$). Non-linear probes bring modest additional gains for most models (Kernel R^2 ranging from 0.85 for VGGish to 0.93 for MERT). An exception is MS-CLAP 2022, where all three probes saturate at $R^2 = 0.78$, indicating that the embedding contains no additional RP information beyond what is linearly encoded.

4.4. Text-based Attribute Prediction

CLAP's joint embedding space enables zero-shot acoustic attribute prediction: the audio-trained linear probe applied directly to text embeddings converts a natural language description into a predicted attribute value. We present this as a qualitative demonstration of cross-modal consistency, as the reduced and hand-crafted prompt set is not designed for statistical generalization.

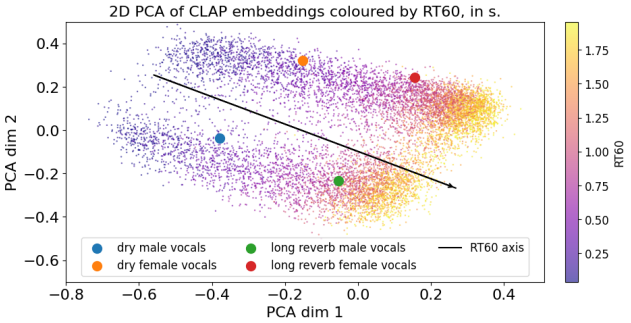


Figure 4: PCA projection of LAION-CLAP embeddings for the VCTK-Corpus test set, colored by RT60. Projected text descriptions confirm that the qualitative descriptors *dry* and *long reverb* land at geometrically coherent positions along the RT60 axis identified by the Linear probe (black arrow).

Figure 4 shows the PCA projection of VCTK-Corpus test-set embeddings, colored by RT60. A monotonically increasing RT60 direction runs from upper left to lower right, consistent with the Linear probe axis (black arrow). Two illustrative descriptors, *dry* and *long reverb*, land at coherent positions along this axis.

Table 4: Predicted RT60 values obtained by applying the VCTK-Corpus linear probe to CLAP text embeddings. Acoustic descriptors (above rule) and semantic controls (below rule) are ordered by expected RT60.

Description	Predicted RT60 (s)
anechoic	0.08
dry	−0.08
slightly reverberant	0.82
moderately reverberant	1.10
highly reverberant	1.80
recording studio	0.09
living room	0.29
concert hall	1.13
cathedral	0.89

The five descriptions in Table 4 form a monotonic sequence from near-zero to 1.80 s. The negative prediction for *dry* (−0.08 s) is an artefact of the unconstrained probe near the lower training boundary ($\mathcal{U}(0.0,\ 2.0)$ s), and its absolute value (0.08 s) is coherent with *anechoic*. *Recording studio* (0.09 s), *living room* (0.29 s),

and *concert hall* (1.13 s) form a monotonically increasing sequence, while *cathedral* yields only 0.89 s, well below the physically expected >2 s, which lies outside the training range $\mathcal{U}(0.0,\ 2.0)$ s.

Repeating the same experiment with LUFS produced no consistent ordering, contrasting with the strong audio-side encoding (Linear $R^2 = 0.92$ on VCTK-Corpus). In addition, a similar demonstration for RP is not straightforward: pitch descriptors in natural language are inherently relational, lacking a fixed absolute reference analogous to reverberation decay time, and the domain-dependence of the RP axis identified in Sec. 4.1 further limits the interpretability of any text-side prediction. Improving audio–text alignment for attributes with relational or domain-dependent language representations remains an open direction for future work.

5. CONCLUSION

We presented a systematic probing study of three fundamental perceptual dimensions, reverberation (RT60), loudness (LUFS), and spectral content (SC and RP), in CLAP audio embeddings, using Linear, MLP, and Kernel Ridge Regression probes across five datasets spanning noise, speech, and music.

Our primary finding is that all the attributes we probed are well represented in CLAP and other audio foundation model embeddings: non-linear probes recover each attribute reliably across all datasets and model families. Within this broader picture, two encoding regimes emerge: RT60, LUFS, and RP are approximately linearly encoded and recoverable with a simple linear probe, while SC requires non-linear probes. For the linearly encoded features, the identified feature-axis directions are geometrically consistent for RT60 and LUFS across datasets, reflecting intrinsic embedding structure. For RP, in contrast, the feature-axis directions have been found to be domain-specific. Both regimes generalize across eight additional audio foundation models, with the exception that amplitude-invariant architectures (Wav2Vec2, WavLM-Large, and MERT) discard loudness entirely by construction. Notably, both MS-CLAP versions and LAION-CLAP are the only models achieving partial linear SC recovery in at least one domain, suggesting that spectral content may become linearly encoded under certain architectural, training conditions, or data regimes.

These findings have direct implications for audio practitioners: the encoding of all three perceptual dimensions in a shared embedding space opens the possibility of exploiting a single frozen foundation model for simultaneous estimation of reverberation, loudness, and spectral content. The geometric consistency of RT60 and LUFS feature axes further hints that such directions could inform workflows such as automatic mix analysis, effect-chain estimation, and text-driven effect control, though validating these applications remains future work. Limitations include the exclusive focus on final-layer embeddings, the use of shoebox room geometry for RT60 augmentation, and residual reverberation in music mixtures (MusDB18HQ, SonicMaster): although the applied RIR dominates the target RT60, these mixes cannot be guaranteed fully dry, so a small non-zero RT60 may be present prior to augmentation. Extending the attribute set (to e.g. tempo, dynamics, and stereo imaging) is a natural next step.

6. REFERENCES

[1] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 336–340.

[3] A. Zhang, E. Thomaz, and L. Lu, "Transformation of audio embeddings into interpretable, concept-based representations," 2025. [Online]. Available: https://arxiv.org/abs/2504.14076

[4] X. Li, W. Chen, Z. Ma, X. Xu, Y. Liang, Z. Zheng, Q. Kong, and X. Chen, "Drcap: Decoding clap latents with retrieval-augmented generation for zero-shot audio captioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[5] S. Deshmukh, D. Alharthi, B. Elizalde, H. Gamper, M. Al Ismail, R. Singh, B. Raj, and H. Wang, "Pam: Prompting audio-language models for audio quality assessment," in *Proc. Interspeech 2024*, 2024, pp. 3320–3324.

[6] S.-W. Chung and M.-S. Choi, "Listen through the sound: Generative speech restoration leveraging acoustic context representation," in *Proc. Interspeech 2025*, 2025, pp. 4843–4847.

[7] A. Chu, P. O'Reilly, J. Barnett, and B. Pardo, "Text2FX: Harnessing clap embeddings for text-guided audio effects," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[8] A. Vosoughi, Y. Zang, Q. Yang, N. Paek, R. Leistikow, and C. Xu, "Multimodal room impulse response generation through latent rectified flow matching," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 14 627–14 631.

[9] K. Kim, J. Koo, S. Lee, H. Joung, and K. Lee, "TokenSynth: A token-based neural synthesizer for instrument cloning and text-to-instrument," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[10] Q. Yang, R. Leistikow, and Y. Zang, "FlowSynth: Instrument generation through distributional flow matching and test-time search," *arXiv preprint arXiv:2510.21667*, 2025.

[11] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-An-Audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 13 916–13 932.

[12] Y. Yuan, Z. Chen, X. Liu, H. Liu, X. Xu, D. Jia, Y. Chen, M. D. Plumbley, and W. Wang, "T-clap: Temporal-enhanced contrastive language-audio pretraining," in *IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2024, pp. 1–6.

[13] K. Seki, Y. Okamoto, K. Yamaoka, Y. Saito, S. Takamichi, and H. Saruwatari, "Spatial-clap: Learning spatially-aware audio–text embeddings for multi-source conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 14 742–14 746.

[14] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *IEEE Automatic*

Speech Recognition and Understanding Workshop (ASRU). IEEE, Dec. 2019, pp. 726–733. [Online]. Available: http://dx.doi.org/10.1109/ASRU46091.2019.9003979

[15] C.-K. Yang, N. Ho, Y.-J. Lee, and H. yi Lee, "Audiolens: A closer look at auditory attribute perception of large audio-language models," 2025. [Online]. Available: https://arxiv.org/abs/2506.05140

[16] W.-C. Chen, C. yu Huang, and H. yi Lee, "Causal tracing of audio-text fusion in large audio language models," 2026. [Online]. Available: https://arxiv.org/abs/2603.13768

[17] T.-Y. Wu, Y.-X. Lin, and T.-W. Weng, "AND: Audio network dissection for interpreting deep acoustic models," in *International Conference on Machine Learning (ICML)*. PMLR, 2024, pp. 53 656–53 680.

[18] Q. Deng, B. Pardo, and T. N. Pappas, "Do joint language-audio embeddings encode perceptual timbre semantics?" *arXiv preprint arXiv:2510.14249*, 2025.

[19] G. Velissaridis, R. Athwal, M. Musharaf, G. Fazekas, and C. Saitis, "Evaluating foundation models on timbre-related cognitive tasks," in *1st Workshop on Large Language Models for Music & Audio (LLM4MA)*, 2025.

[20] V. Deng, C. Wang, G. Richard, and B. McFee, "Investigating the sensitivity of pre-trained audio embeddings to common effects," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[21] Y.-T. Yeh, J. Koo, M. Martínez-Ramírez, W.-H. Liao, Y.-H. Yang, and Y. Mitsufuji, "Fx-encoder++: Extracting instrument-wise audio effects representations from mixtures," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025.

[22] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pre-training with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[23] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[24] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[25] ITU-R, "Algorithms to measure audio programme loudness and true-peak audio level," International Telecommunication Union, Recommendation ITU-R BS.1770-5, 11 2023. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1770-5-202311-I/en

[26] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015, vol. 5.

[27] J. O. S. III, *Spectral Audio Signal Processing*. W3K Publishing, 2011. [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/

[28] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[29] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 1068–1077.

[30] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[31] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of musdb18," Dec. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[32] J. Melechovsky, A. Mehrish, A. Roy, and D. Herremans, "SonicMaster: Towards controllable all-in-one music restoration and mastering," *arXiv preprint arXiv:2508.03448*, 2025.

[33] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[34] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," in *International Conference on Learning Representations (ICLR)*, 2024, pp. 12 181–12 204.

[35] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12 449–12 460, 2020.

[36] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 28 492–28 518.

[38] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[39] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting Fréchet audio distance for generative music evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1331–1335.