Learning Linearity in Audio Consistency Autoencoders via Implicit Regularization

Bernardo Torres[1] Manuel Moussallam[2] Gabriel Meseguer-Brocal[2]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris [2]Deezer Research

Audio demos & code ↓

Models: `pip install lin-cae`

Key Contributions

- Method to train consistency autoencoders that **behave linearly** with **no extra loss terms and no architecture change**.
- Lin-CAE **matches reconstruction quality** of non-linear baselines.
- Emergent linear properties enable **source separation via plain latent arithmetic**.

1 Audio autoencoders are not linear

Perception is not linear — but sound waves are.
Audio obeys **superposition**: mix two signals, get their sum. In images, pixel addition is physically meaningless.

Neural audio autoencoders don't inherit this property:

- They compress as much as possible at the expense of latent space structure, or
- They are designed for semantic manipulation, not signal-level processing

A **linear** latent space enables direct audio composition and volume control in the compressed space, avoiding the need for expensive cycles of encoding/decoding.

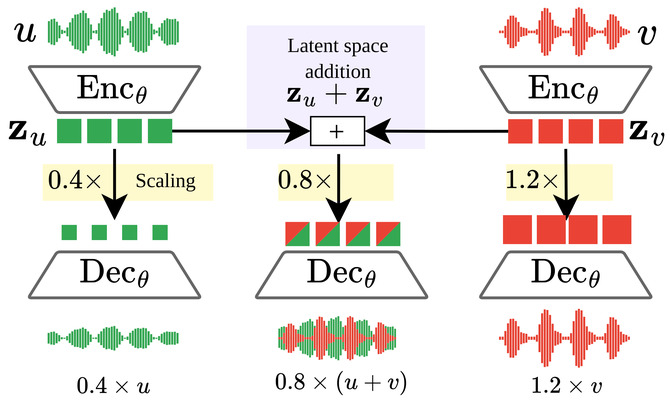

Figure 1: Illustration of a linear decoder

2 Consistency Autoencoders (CAE)

We build on **Music2Latent** (Pasini et al, 2024):

- **64× compression** at 44.1 kHz
- **Generative (diffusion) decoder**, with **single-step decoding**
- The proposed method, however, is model-agnostic

3 Properties enforced during training

Homogeneity — equivariance to scalar gain:

$$\mathrm{Dec}(a \cdot z_x) \approx a \cdot \mathrm{Dec}(z_x)$$

Additivity — latent sum equals sum in signal space:

$$\mathrm{Dec}(z_u + z_v) \approx \mathrm{Dec}(z_u) + \mathrm{Dec}(z_v)$$

4 Data augmentation and latent space tricks

No extra loss terms, no architecture changes, same database. Only two tricks applied to the decoder:

- Trick 1: **Infer scale** from latent magnitude
- Trick 2: **Map summed latents** to an audio mixture

For this, we create **mixtures of mixtures**:

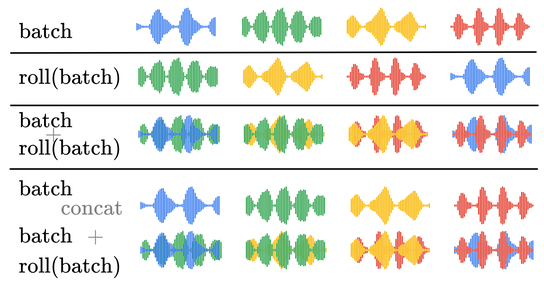

Figure 2: Artificial mixes are created on-the-fly by rolling the batch and adding. Half of the batch are originals, half are mixtures of mixtures.

5 Algorithm (intuition)

Require: Dataset $\mathcal{D}$, Encoder E, Decoder D

for each training step **do**
- Sample two audio signals $\boldsymbol{u}, \boldsymbol{v} \sim \mathcal{D}$
- Create mixture of mixtures $\boldsymbol{x} \leftarrow \boldsymbol{u} + \boldsymbol{v}$
- Compute latents: $\boldsymbol{z}_u \leftarrow E(\boldsymbol{u}), \quad \boldsymbol{z}_v \leftarrow E(\boldsymbol{v})$
- **Sum** latents: $\boldsymbol{z}' \leftarrow \boldsymbol{z}_u + \boldsymbol{z}_v$
- Sample **scalar gain**: $\boldsymbol{a} \sim \mathcal{U}(a_{\min}, a_{\max})$
- Decode **scaled** latent **sum**: $\hat{\boldsymbol{x}} \leftarrow D(\boldsymbol{a} \cdot \boldsymbol{z}')$
- Loss: $\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{a} \cdot \boldsymbol{x})$

end for

6 Measuring linearity

Lin-CAE (our model) = same architecture and training as Music2Latent + the augmentation procedure described in Section 4.

MSS = Multi-Scale Spectral distance
Check paper for SDR values

Model	Recons. MSS ↓	Decoder Hom. MSS ↓	Decoder Add. MSS† ↓	Encoder Hom. Err. ↓	Encoder Add. Err.† ↓
Music2Latent-pub*	1.14	2.52	5.01	12.13	2.82
Stable Audio VAE	0.72	3.03	5.38	4.59	1.71
Music2Latent-ours**	0.98	2.27	5.21	8.52	2.73
Lin-CAE	1.01	**1.37**	**0.99**	**0.69**	**0.60**

* Public weights. ** Retrained on the same data as Lin-CAE. †Additivity measured on MUSDB18-HQ.

- Lin-CAE shows both homogeneity and additivity while **preserving reconstruction quality.**
- **Additivity generalizes to 4 sources** (measured on MUSDB18-HQ)
- Emergence in encoder linearity

7 Oracle source separation via latent subtraction

Despite trained on positive gains, decoder generalizes to subtraction:

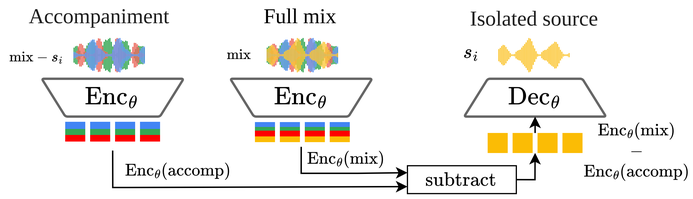

Figure 3: Estimate a source by subtracting the accompaniment latent from the mix latent.

Model	Recons. MSS ↓	Separation MSS ↓ Bass	Drums	Other	Vocals
Music2Latent-pub*	1.16	95.34	28.08	6.78	5.10
Stable Audio VAE	0.57	323.26	87.80	3.65	3.08
Music2Latent-ours**	0.97	39.56	12.26	5.30	3.49
Lin-CAE	1.00	**1.58**	**1.16**	**1.21**	**1.23**

* Public weights. ** Retrained on the same data as Lin-CAE.

- **Lin-CAE separation matches its own reconstruction**, while baselines fail by orders of magnitude.
- Listen to the demos! (QR code in the header)

research@deezer.com bernardo@bernardotorres.net

IEEE ICASSP 2026, Barcelona

I'm currently on the job market! — scan for my CV or visit bernardotorres.net