

Dual-Contrastive Sparse Autoencoders Reveal Features of Musical Interpretation

Jan Chen^{1,2}

Manuel Cherep²

Pattie Maes²

Nikhil Singh³

¹Technical University of Munich

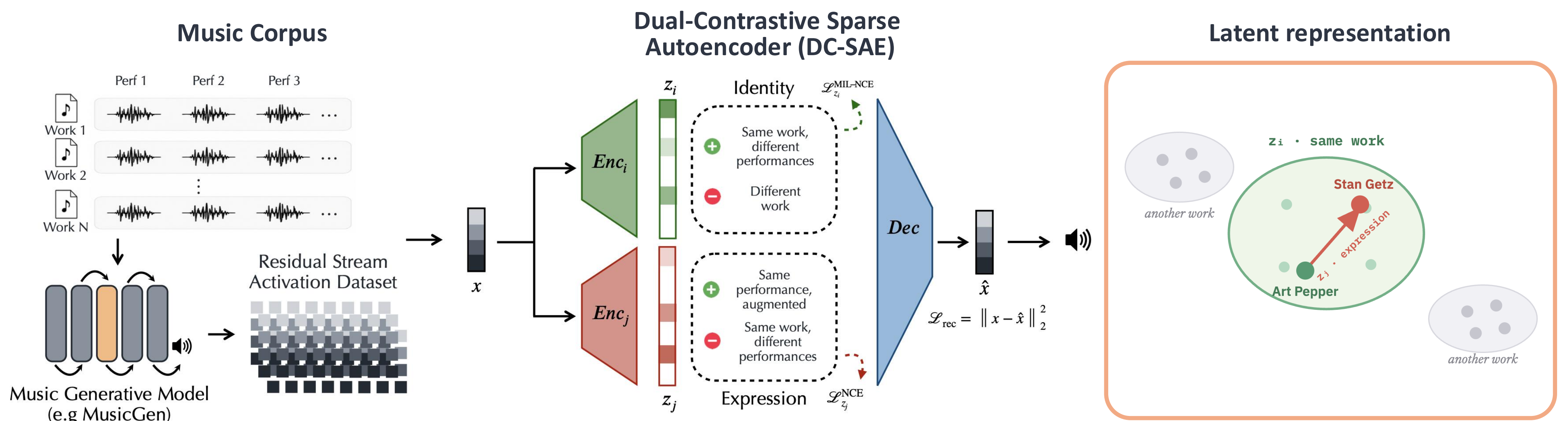
²Massachusetts Institute of Technology

³Dartmouth College



Summary: Which features of a music performance are constitutive of the work, and which express its interpretation? Audio embedding models conflate work identity with performance style, while existing interpretability tools recover flat dictionaries that mix the two. We introduce DC-SAE, which factors MusicGen’s residual stream into interpretable and steerable sparse feature subspaces.

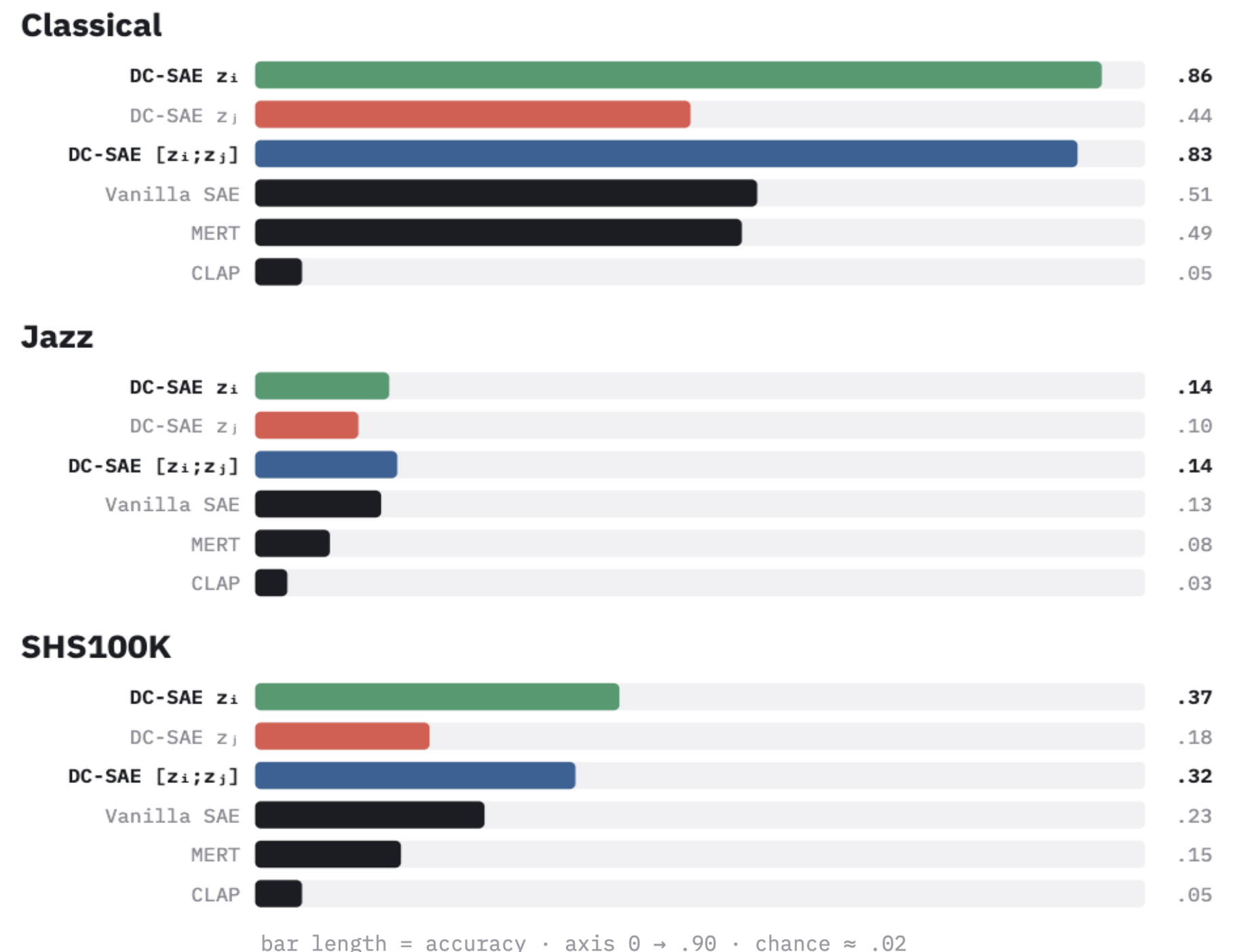
1 Pipeline



2 Evaluation

Work-classification accuracy

We run linear probes based on held-out recordings of musical works, and ask for a prediction of the underlying work given a 30s music. The work-identity branch z_i and the concatenation $[z_i; z_j]$ show a specialization in classifying work.

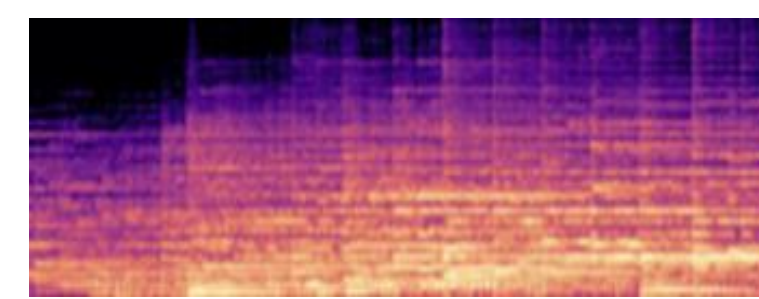


Performer-classification accuracy

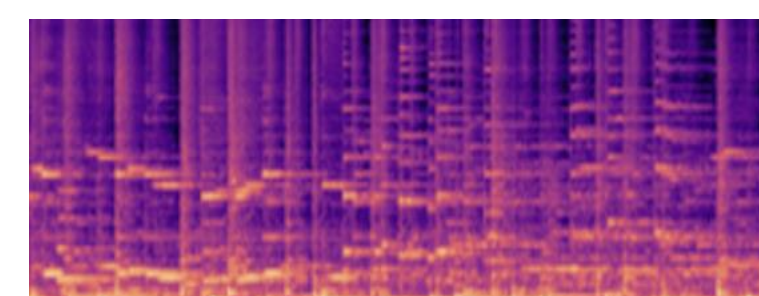
Given 30s of music, we run a linear probe classifying the underlying performer. The expression branch z_j and the concatenation $[z_i; z_j]$ show an emergent behavior to classify



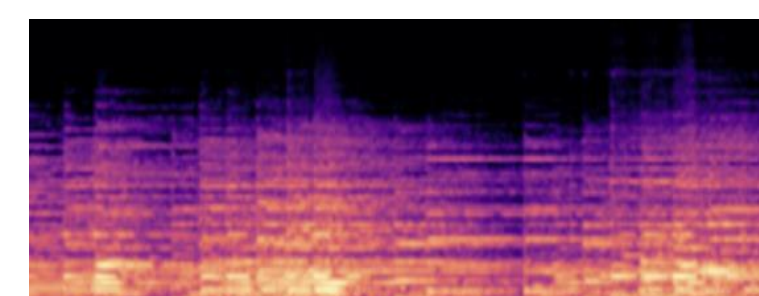
3 Examples of Expression Features



£59100: *expressivo/rubato (piano)*. Solo piano with expressive phrasing and rubato, romantic-leaning, especially Chopin.



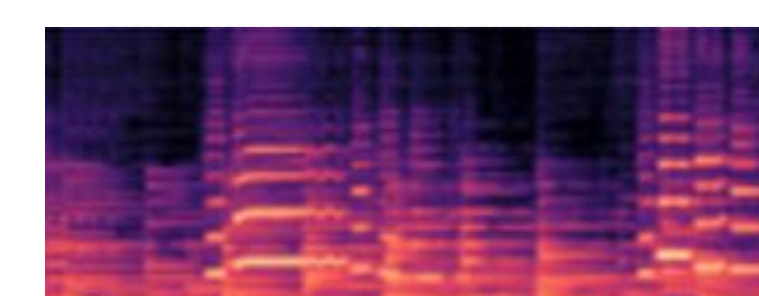
£3876: *jazz piano improv.* Fast jazz piano solo lines.



£40006: *vibrato-violin / wave-like piano*. Vibrato on classical pieces for violin. Wave-like fast arpeggiated chord textures on piano.

4 Interpretation and Steering

1 Two performances of the same work



source · Stan Getz



target · Art Pepper

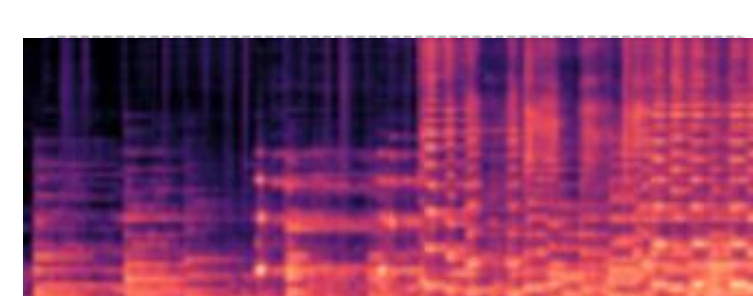
2 Rank z_j features by target – source contrast

which single feature best explains the difference between the two interpretations

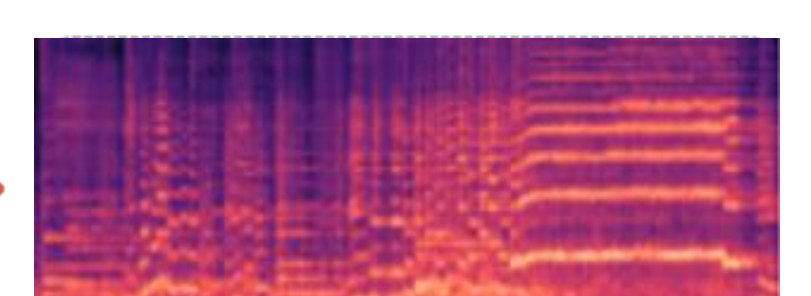


£57490 top-1 direction: human evaluation confirms it is interpretable as a prominent high-pitched saxophone

3 Inject it into a neutral prompt



prompt “jazz saxophone music”



follows Art Pepper

We run linear probes based on held-out recordings of musical works, and ask for a prediction of the underlying work given a 30s music. The work-identity branch z_i and the concatenation $[z_i; z_j]$ show a specialization in classifying work.