Sparse Autoencoders for Interpretable Emotion Control in Text-to-Speech

Hongfei Du Jiacheng Shi Sidi Lu Gang Zhou Ye Gao

Department of Computer Science, William & Mary, USA

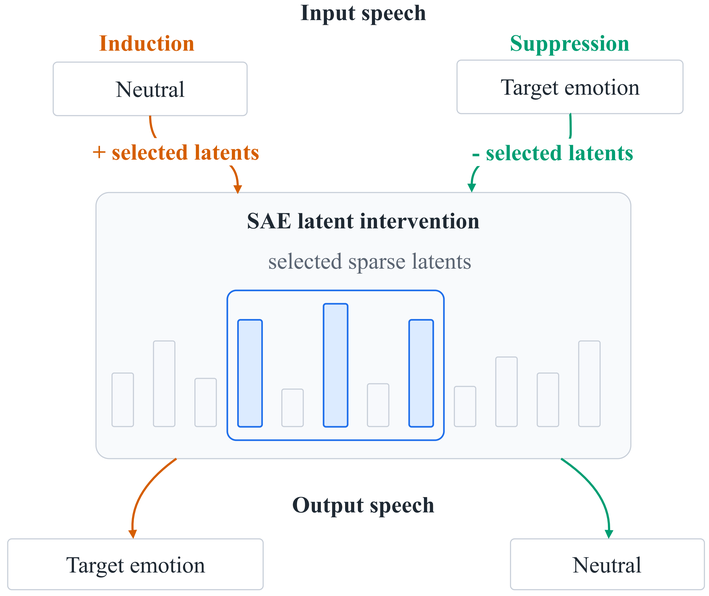
Audio Demo

Motivation

- LLM-based TTS improves speech expressiveness, making controllable emotional expression increasingly important.
- Existing label- and reference-based methods rely on external conditioning, offering limited interpretability and flexibility once input conditions are fixed.
- Dense activation steering enables training-free emotion control, but captures emotion as a single global representation shift rather than separable feature-level variation.
- We use SAEs to decompose emotion-related variation in the TTS semantic backbone into sparse latent features for interpretable bidirectional control.

Conceptual View

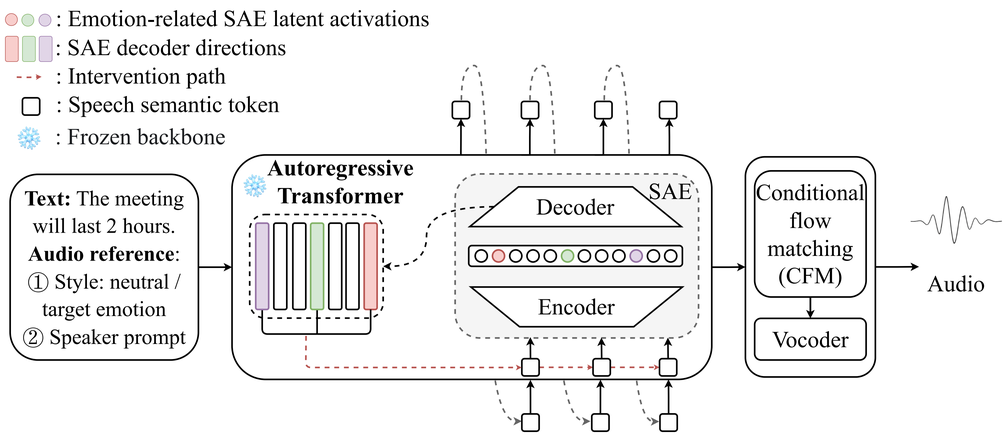
Same text content and speaker identity held fixed



Sparse Autoencoders

- Represents each hidden state with only a small number of active latent features.
- Transforms dense, entangled activations into sparse features that are easier to interpret.
- Provides a feature-level interface for identifying and steering emotion-related variation in the TTS semantic backbone.

Approach Overview



1. Identify emotion-related latents

Rank SAE features by sentence-level emotion selectivity: $\Delta_i^{(e)} = \mathbb{E}_u[\mathbf{1}_i^{(e)}(u) - \mathbf{1}_i^{(\text{neutral})}(u)]$

Select the top-m emotion-related latents; main experiments use m = 6.

2. Feature-level intervention

For selected emotion latents $j \in \mathcal{F}_e$, shift their activations: $a_j^{\text{new}}(x_{l,t}) = a_j(x_{l,t}) + \alpha_e$

Positive α_e induces the target emotion; negative α_e suppresses it.

Experiment Setup

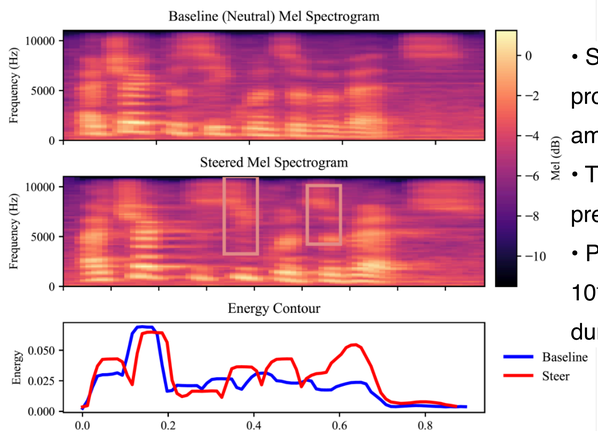
SAE training and controlled data

- Top-k SAE trained on layer-16 semantic-backbone residual activations
- 56,000 emotion-controlled TTS generations
- 4,096 SAE latents; k = 32 active features per token
- 400 texts × 7 emotions × 20 speaker-timbre references

Controlled steering evaluation

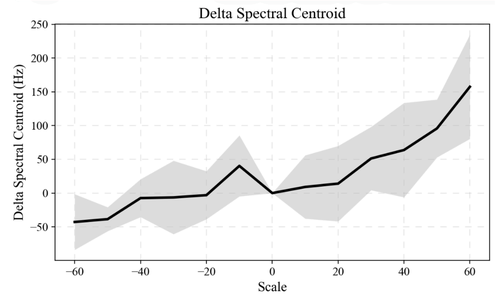
- Matched text-speaker pairs; only emotional style varies
- Selectivity analysis: 43,408 matched generations
- Steering: 100 matched cases per target emotion
- Targets: anger, happiness, sadness; induction and suppression
- Metrics: Emo-SIM, WER, Spk-SIM, human EMOS/NMOS

Pitch/Energy Validation



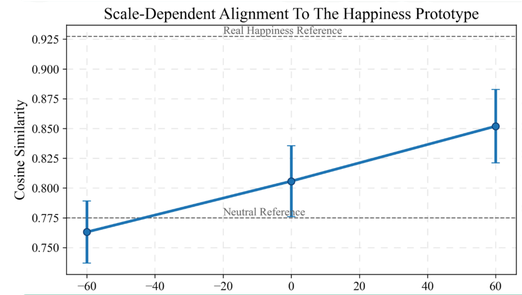
- Steering one emotion-related latent produces localized mid/high-frequency amplification.
- The overall time-frequency structure is preserved.
- Paired analysis: F0 +23.11 Hz ($p = 1.07 \times 10^{-4}$), RMS energy +0.00435 ($p = 0.00769$), duration n.s. ($p = 0.687$).

Brightness Validation



Positive scales increase spectral centroid; negative scales lower it.

Continuous Intensity Control



A single steering coefficient provides continuous control toward the target-emotion prototype.

Bidirectional Steering

Emotion induction: Neutral→Target

Method	Anger Emo↑	Anger WER↓	Anger Spk↑	Happiness Emo↑	Happiness WER↓	Happiness Spk↑	Sadness Emo↑	Sadness WER↓	Sadness Spk↑
VALL-E-X	0.831	3.1	0.302	0.697	5.3	0.320	0.869	7.8	0.352
Spark-TTS	0.857	2.7	0.488	0.770	8.6	0.463	**0.907**	2.3	0.523
EmoVoice	0.806	4.1	0.358	0.728	3.4	0.342	0.850	4.0	0.386
CosyVoice	0.813	3.9	0.569	0.712	2.9	**0.597**	0.799	2.4	0.605
Random SAE	0.892	1.4	**0.628**	0.813	6.0	0.461	0.858	1.7	**0.637**
Global	0.910	**0.1**	0.552	0.879	4.0	0.495	0.876	1.9	0.516
SAE-Emotion	**0.912**	0.3	0.569	**0.885**	**2.2**	0.515	0.880	**1.5**	0.481

Emotion Supression: Target→Neutral

Method	Anger Emo↑	Anger WER↓	Anger Spk↑	Happiness Emo↑	Happiness WER↓	Happiness Spk↑	Sadness Emo↑	Sadness WER↓	Sadness Spk↑
Random SAE	0.841	**0.8**	0.342	0.886	2.14	0.343	0.939	**0.77**	0.427
Global	0.915	2.6	**0.392**	0.920	**1.48**	**0.379**	0.933	1.63	0.436
SAE-Emotion	**0.939**	2.8	0.374	**0.924**	2.31	0.301	**0.941**	0.80	**0.441**

Perceptual Quality

Method	EMOS	NMOS
SAE-Emotion	**3.22**	**3.49**
Global Steering	3.10	3.38
Random SAE	1.82	3.22

Randomized blind listening study with 20 raters; 0–5 scale,

Takeaway

- Emotional variation in the TTS semantic backbone is distributed across multiple sparse SAE latents.
- A small subset of emotion-related latents shows distinct activation and acoustic patterns under controlled conditions.
- Selectivity-ranked latent intervention enables bidirectional emotion control without backbone updates, while largely preserving content, naturalness, and speaker identity.