What Really Carries Emotion in TTS?

Zonos EDA, Kipling matching, and artifact analysis

Kacper Wikiel

IYPT-style acoustic research

May 2026

Research Question

Question

Does emotional TTS encode emotion mainly through pitch, tempo, energy, pauses, or timbre?

Original claim tested here

- Emotional control may not be a full actor-like performance.
- A model may simulate emotion by moving a few acoustic levers.
- We need measured audio features, not definitions or vendor claims.

Hypotheses

1. **Simple-channel hypothesis**: intended emotion is decodable from a small subset of acoustic channels.
2. **Nonlinearity hypothesis**: increasing an emotion vector is not a linear amplification of one feature.
3. **Channel mismatch hypothesis**: the channel that moves most physically is not necessarily the channel that best classifies emotion.

What would make this interesting?

If timbre changes most, but pitch alone classifies emotion best, the model is not just doing one obvious prosody trick.

Actual Experiment Run

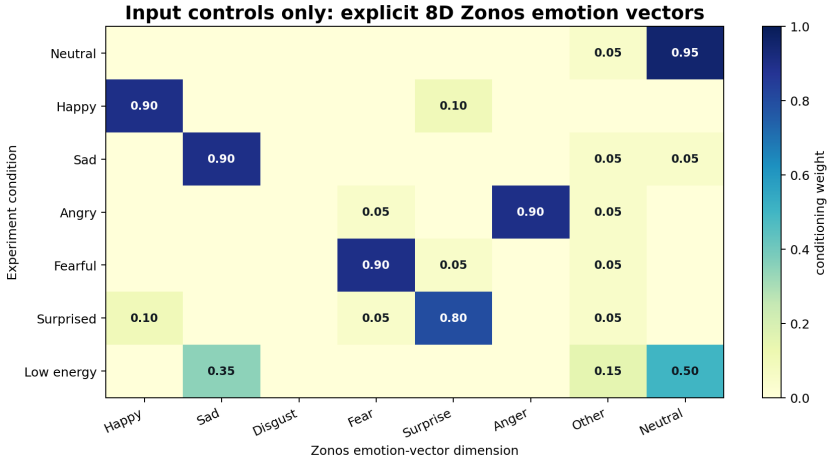
Generated dataset

- Model: Zonos v0.1 transformer
- 5 fixed English sentences
- 7 conditions: neutral, happy, sad, angry, fearful, surprised, low energy
- 35 base WAV files
- 12 additional sweep WAV files

Controls

- text held fixed
- Zonos emotion vector changed
- pitch_std fixed at 35
- speaking_rate fixed at 15
- CFG fixed at 2.0
- short-sentence EDA used Zonos' default speaker path
- later poem matching fixes speaker identity explicitly

Input Control, Not a Correlation Matrix



Rows are experiment conditions. Columns are Zonos emotion-vector dimensions. The result matrix starts later.

Feature Extraction

Feature map: each metric is a physical hypothesis

The analysis asks which acoustic channel moves when the model is told to synthesize an emotion.

Pitch

F0 mean
F0 std
F0 range
Voiced ratio

Melody and intonation.
Main candidate for happy/angry/fear.

Tempo

Duration
Words/s
Syllables/s

How quickly speech unfolds.
Sad and low-energy speech should slow down.

Energy

RMS mean
RMS std

Loudness and emphasis.
Anger may raise mean or variance.

Pauses

Pause count
Pause total
Pause max

Silence and hesitation.
May separate sad/fearful styles.

Timbre

Spectral centroid
MFCC mean/std

Voice color and spectral envelope.
Whisper/strain should appear here.

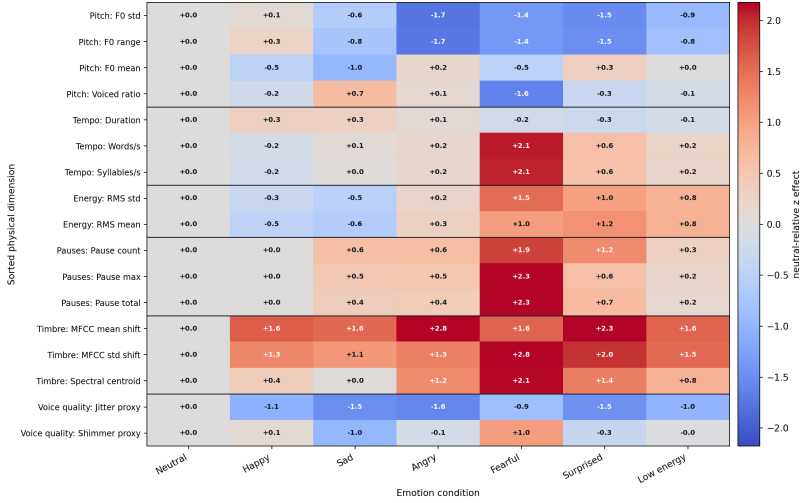
Voice quality

Jitter proxy
Shimmer proxy

Frame-to-frame instability.
Detects roughness or unstable voicing.

EDA: Emotions on X, Physical Dimensions on Y

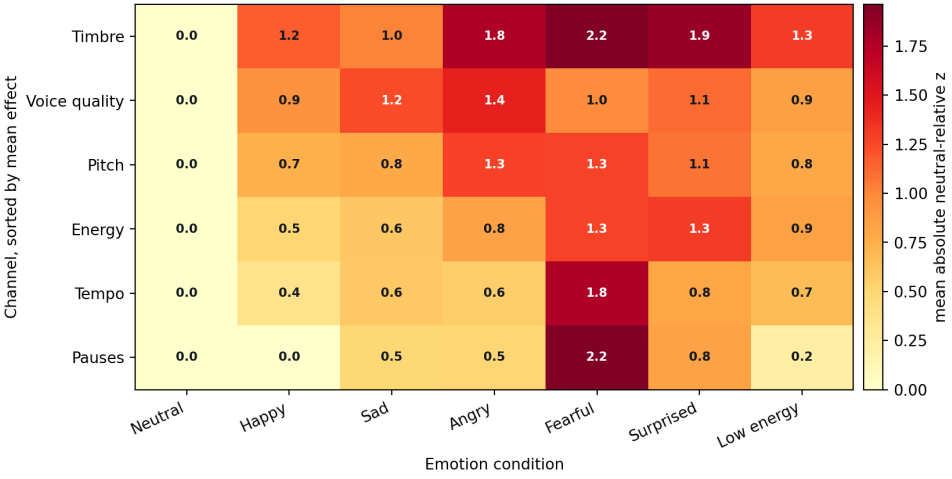
Cell value = mean normalized change relative to the neutral output of the same sentence.



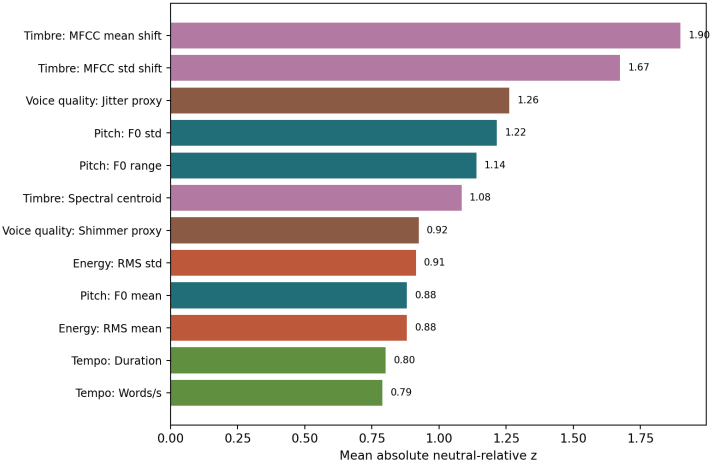
First Read of the EDA

- Fearful is the strongest global perturbation: pauses, tempo, energy, timbre, and centroid all move.
- Happy is not simply “higher pitch”: its F0 mean goes down on average, while MFCC shift is large.
- Angry, fearful, and surprised reduce F0 std/range relative to neutral in this run.
- Timbre moves for nearly every non-neutral condition, so emotion is not only speed or pitch.
- Neutral is zero by construction because every sample is compared against its same-sentence neutral baseline.

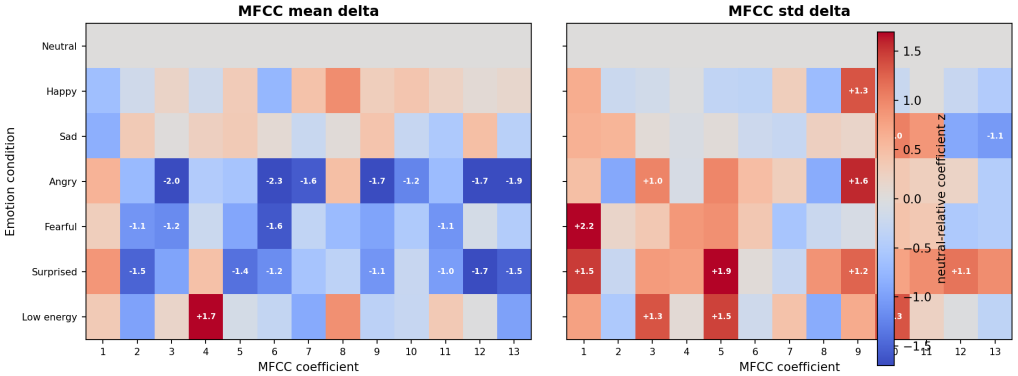
Which Channel Moves Most?



Top Moving Physical Dimensions

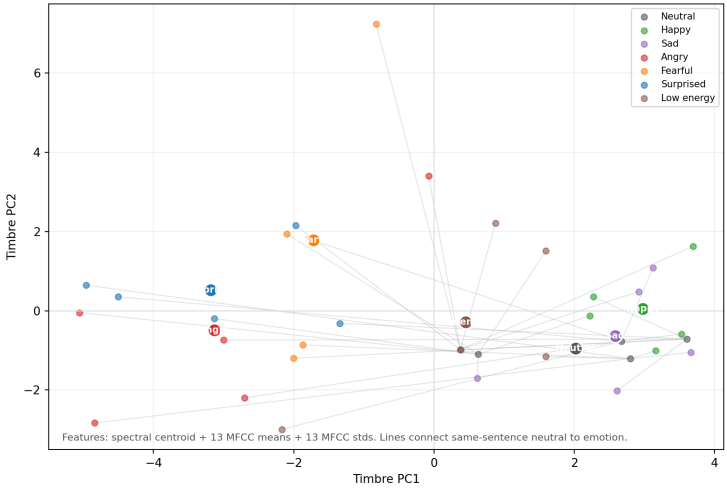


Timbre Fingerprint: MFCC Changes



MFCC coefficients are a compact spectral-envelope vector: a better timbre object than raw spectrogram pixels.

Timbre Space



How to Read Timbre Visuals

- MFCC coefficients summarize the spectral envelope, so they are closer to barwa than raw waveform energy.
- The MFCC heatmap is a fixed-size vector view: it can be averaged, tested with ANOVA, and fed to a classifier.
- The PCA plot uses only timbre features, so any separation there is not caused directly by pitch or speech rate.
- This is stronger evidence than spectrogram subtraction because the axes are stable physical descriptors.

Statistical Results

Test	Strongest features	Score	Interpretation
ANOVA	F0 std	F=9.01, p=1.62e-05	Pitch variability differs strongly between emotion conditions.
ANOVA	MFCC std shift	F=8.39, p=2.97e-05	Timbre distribution changes are significant.
ANOVA	MFCC mean shift	F=7.56, p=6.94e-05	Voice color shifts across conditions.
ANOVA	Pause total	F=5.70, p=5.70e-04	Some emotions insert timing structure.
Mutual information	F0 std, F0 range	MI=0.635, 0.582	Pitch variation carries the most label information.
Mutual information	MFCC std, MFCC mean	MI=0.521, 0.467	Timbre is the second strongest signal.

Channel Ablation Classifier

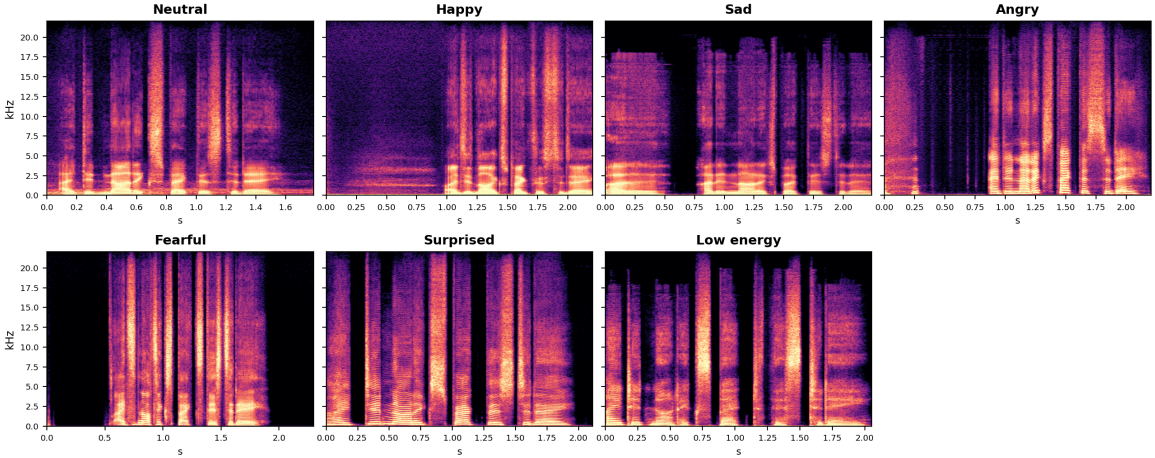
Leave-one-sentence-out 7-class classification

Random baseline is about 0.14. A channel is useful if it predicts emotion on held-out text.

Feature channel	Accuracy
Pitch only	**0.60**
Timbre only	0.37
Voice quality only	0.26
Pauses only	0.23
Energy only	0.20
Tempo only	0.17

Pitch is the best emotion label decoder, even though timbre has the largest physical movement.

Spectrograms: One Fixed Sentence



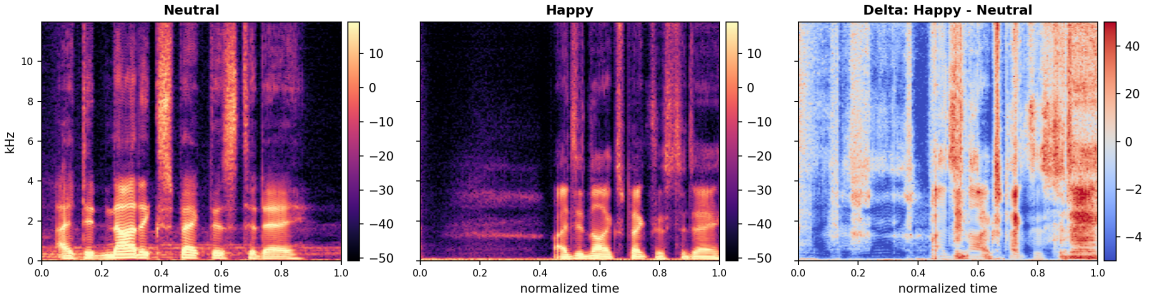
Spectrogram Observations

- Fearful contains a long silence block before speech, matching the pause spike in the heatmap.
- Angry and surprised show stronger high-frequency structure than neutral.
- Happy changes the time envelope and high-frequency texture, not just pitch.
- Low energy is not simply quieter; it still shows strong vertical energy bursts.

Interpretation

The emotion vector changes the temporal and spectral shape of the utterance. The result is richer than a pure speed knob, but still measurable through simple acoustic channels.

Delta Spectrogram: Visual Sanity Check



Naive visual subtraction after time-normalization. Useful as a sanity check, not as the main metric.

This is useful to see that something changed, but it is not the main evidence.

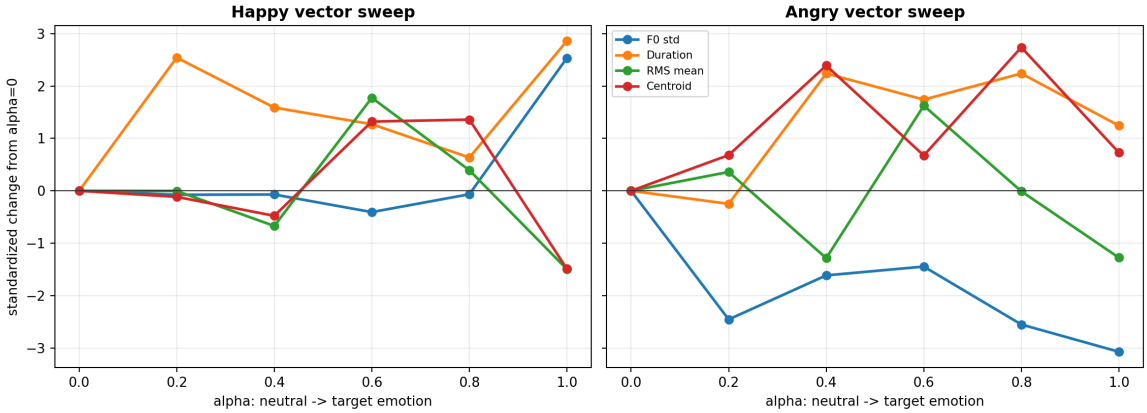
Why Spectrogram Delta Is Not the Metric

Skeptical point

Subtracting spectrogram images is visually intuitive, but it does not scale cleanly into a physical conclusion.

- Emotional TTS changes duration, pauses, and phoneme timing, so frames are not naturally aligned.
- A large delta may mean different timing, not different timbre.
- The image is high-dimensional and hard to reduce to a stable vector across many samples.
- Therefore the deck uses MFCC, spectral centroid, F0, RMS, pause, and rate features as the quantitative evidence.

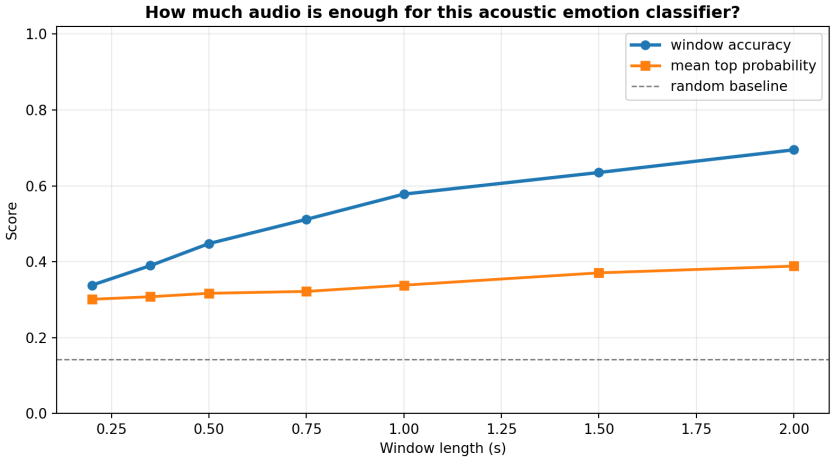
Nonlinearity Probe



Nonlinearity Result

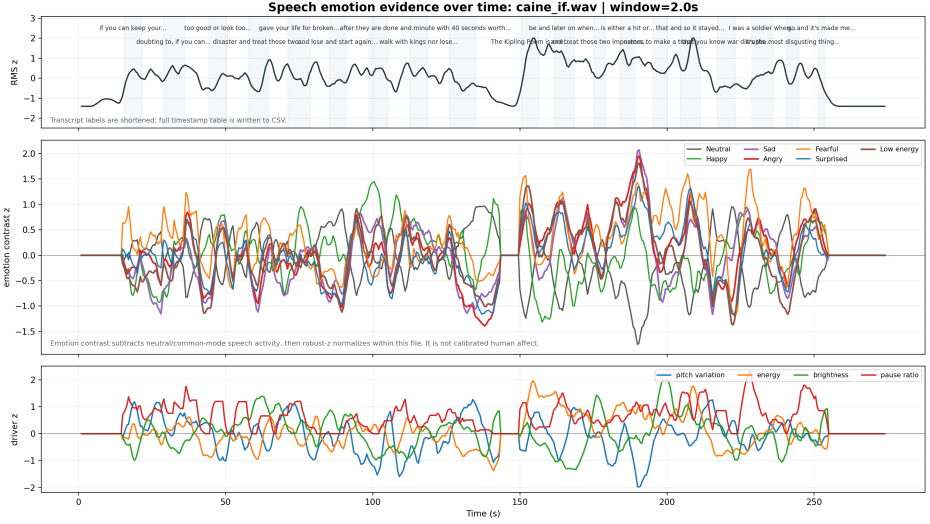
- The happy sweep is not monotonic: duration jumps early, F0 std changes mostly near the endpoint.
- The angry sweep is also not a straight line: centroid and duration peak at intermediate settings.
- Holding the random seed fixed across alpha values reduces the sampling confound.
- This supports the nonlinearity hypothesis: the emotion vector changes model state, not one scalar acoustic knob.

How Long Is Enough?

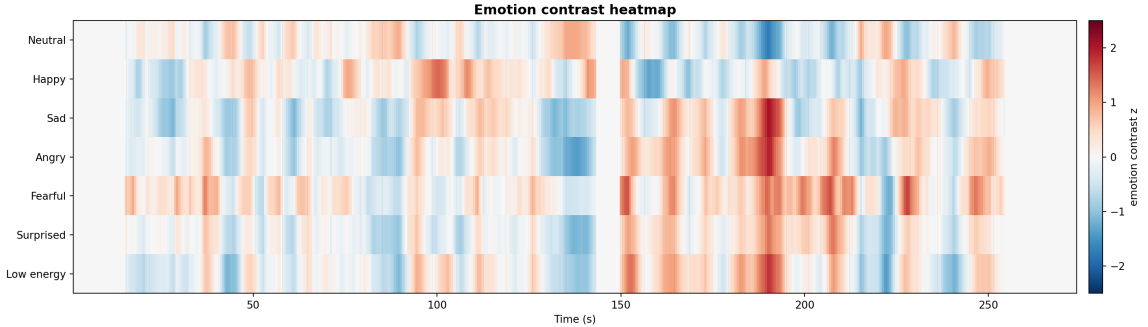


In this local acoustic classifier, 1.0 s is usable but weak; 1.5–2.0 s is more defensible for a timeline.

Real Speech Check: Emotion Over Time



Real Speech Check: Contrast Heatmap



Contrast subtracts neutral/common speech activity, so the plot shows what increases locally, not just loudness.

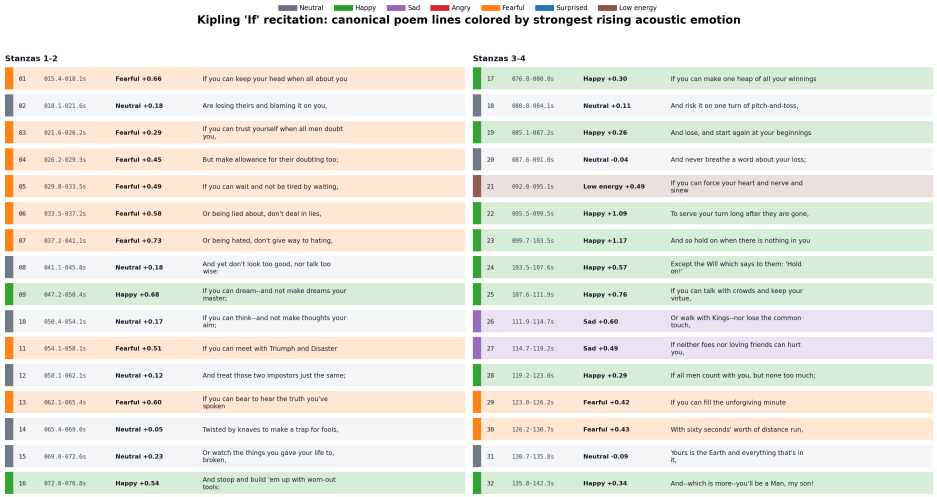
From EDA to Matching

New task

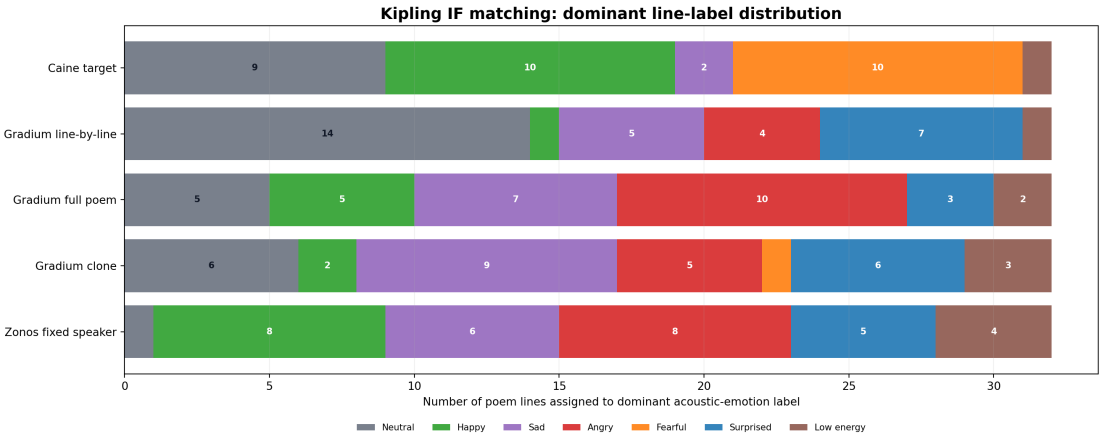
Use the Michael Caine recitation of Kipling's *If* as a target acoustic-emotion trajectory, then test whether controllable TTS can reproduce that trajectory.

- Target is not a single emotion label for the whole poem.
- Each canonical poem line gets a vector of evidence: emotion contrasts plus acoustic drivers.
- Dominant emotion is only a compressed summary used for quick comparison.
- The real research question becomes: can TTS follow a time-varying acoustic performance?

Caine Target: Canonical Poem Lines



Caine Target Distribution



Caine target is mostly fearful, happy, and neutral. Most TTS attempts drift into other acoustic regions.

Gradium: Useful Negative Control

What Gradium controls

- voice choice and custom voice cloning
- speed via padding controls
- sampling temperature
- CFG / voice similarity
- text and pause tags

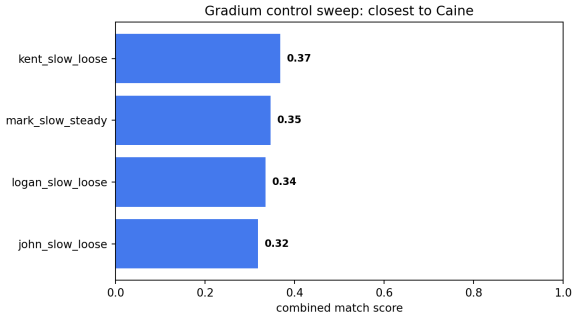
What it did not give us

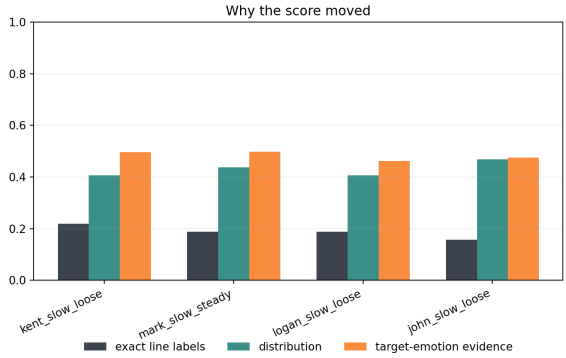
- no direct emotion-vector trajectory
- cloning changed speaker/timbre, not performance arc
- full-poem generation beat line-by-line prosody
- but still did not converge to Caine's line map

Result

Gradium is a good voice/timbre baseline, but not the right primary model for causal emotion-control research.

Gradium Sweep Results





Duration was removed from the score; the comparison is emotion-pattern only.

Zonos Matching Design

Why Zonos

- explicit 8D emotion vector
- direct fear, happiness, sadness, anger, surprise, neutral axes
- pitch variation and speaking rate are separate controls

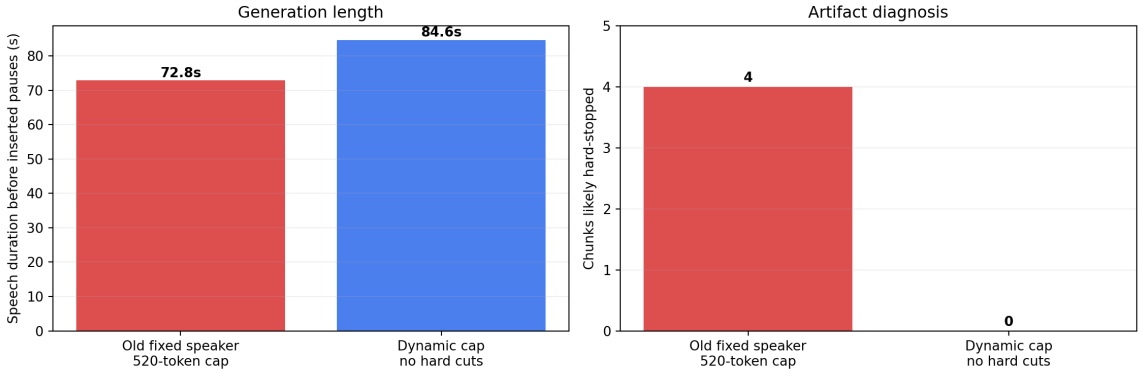
Poem generation

- target = Caine line labels
- consecutive same-label lines merged into chunks
- fixed Kent speaker embedding
- dynamic token cap to avoid clipped chunks

Important correction

Without a fixed speaker embedding, Zonos sounded like multiple people. Emotion matching is meaningless unless identity is controlled.

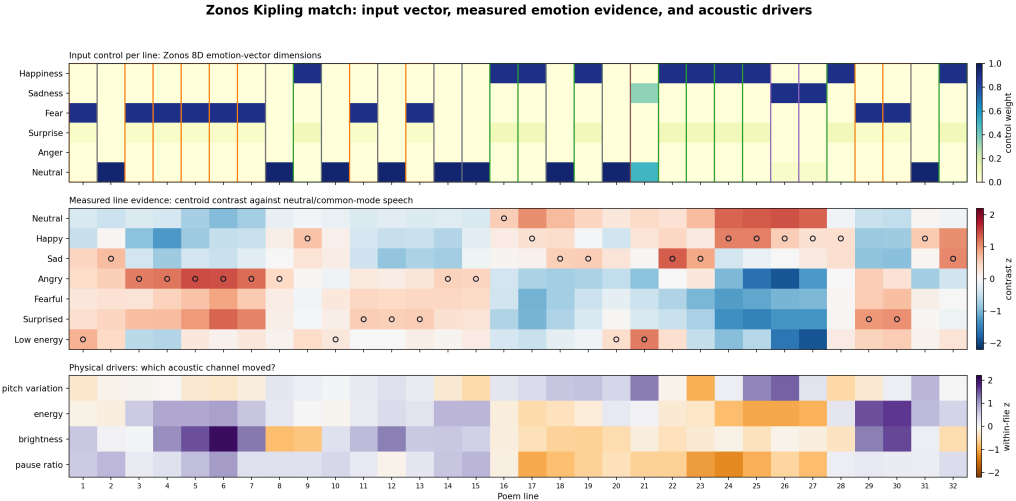
Artifact Diagnosis



The first fixed-speaker Zonos poem had hard stops because `max_new_tokens=520` capped long chunks at about 6 seconds.

Zonos Fixed Speaker: Dominant Line View

Zonos Fixed Speaker: Multidimensional View



Rows 1-6 are the intended Zonos controls. Rows below are measured from audio, so disagreement reveals entanglement/artifacts rather than a label bug.

Why Dominant Emotion Is Not Enough

- A line can have rising fearful, angry, and low-energy evidence at the same time.
- Dominant label discards the second and third strongest axes.
- Physical drivers can disagree with the emotional label: pitch variation, energy, brightness, and pause ratio move separately.
- The multidimensional plot is therefore the real evidence view; the colored transcript is only the readable summary.

Research implication

TTS emotion is not a one-hot state. It is a vector-valued acoustic transformation, and artifacts can masquerade as emotion if not audited.

Matching Results So Far

Run	Duration	Exact	Rank	Dist.	Main observation
Gradium line-by-line	140.8 s	0.219	0.480	0.406	voice/settings sweep; no fear channel reproduced
Gradium full poem	134.4 s	–	–	–	better continuity, but drifts to angry/sad/happy
Gradium clone	138.7 s	0.125	0.337	0.375	cloning changes timbre, not Caine-like performance
Zonos fixed speaker, clipped	77.5 s	0.219	0.449	0.563	invalid: several chunks hard-stopped near 6 s
Zonos fixed speaker, dynamic cap	89.3 s	0.188	0.428	0.375	valid audio; still not converging to Caine

Exact = same dominant line label. Rank = target label ranked high even if not dominant. Dist. = label-distribution similarity.

Interpretation of the Matching Failure

- Explicit emotion vectors do move Zonos acoustics, but not like a human actor following the poem.
- Stronger generation hygiene made the audio more valid, but the match score dropped.
- That is not a failure of the analysis: it means previous artifacts were inflating the apparent match.
- Caine's performance contains continuity, rhetoric, breath, emphasis, and line-to-line intent, not just per-line emotion labels.

Updated thesis

Current TTS controls can steer acoustic channels, but matching a real actor requires controlling a trajectory of multiple physical dimensions, not only selecting emotion labels.

Answer to the Research Question

Measured answer

In the short-sentence Zonos EDA, emotion is carried most visibly by timbre, but most decodably by pitch variation. In the Kipling matching task, this is insufficient: real performance is a multidimensional trajectory.

- Largest physical movement: **timbre** (mean absolute neutral-relative $z = 1.55$).
- Best single-channel classifier: **pitch** (leave-one-sentence-out accuracy $= 0.60$).
- Strongest individual statistical feature: **F0 std** (ANOVA $p = 1.62\text{e-}05$, $MI = 0.635$).
- Matching result: direct emotion vectors alone do not reproduce Caine's poem trajectory.

What Is Still Missing

- Repeat with multiple speaker embeddings to separate emotion from speaker-style drift.
- Add ASR WER to test whether stronger emotion damages intelligibility.
- Add speaker embedding similarity to test whether emotional control changes identity.
- Add parameter sweeps over Zonos emotion intensity, pitch_std, speaking_rate, and CFG.
- Compare against IndexTTS-2, Chatterbox, OmniVoice, FastPitch, and FastSpeech 2.
- Add human ratings: perceived emotion, naturalness, and acting quality.

CodeSOTA Improvement

Gap

A TTS leaderboard should not only rank naturalness. It should expose controllability: emotion, pitch, tempo, energy, pauses, timbre, reference audio, and tags.

- Add structured model fields for each control surface.
- Add a small reproducible benchmark: same sentences, same controls, same feature extractor.
- Report both “what moves” and “what classifies” because those are different claims.

Takeaway

Defensible thesis

Emotional TTS in this experiment is not a full hidden theatrical performance. It is a structured acoustic transformation: strong timbre movement plus pitch features that make the intended label easy to decode.

The research value is the mismatch: the biggest physical effect is timbre, while the strongest classifier channel is pitch.

Artifacts and Sources

- Generated data: `runs/zonos_emotion_research/`
- Kipling matching: `runs/zonos_kipling_match/`
- Gradium controls/cloning: `runs/gradium_caine_match/`
- Experiment script: `scripts/zonos_emotion_research.py`
- Matching script: `scripts/zonos_kipling_match.py`
- Slides: `slides/emotion_tts_iyp.pdf`
- Zonos: https://github.com/Zyphra/Zonos
- Zonos conditioning: https://github.com/Zyphra/Zonos/blob/main/CONDITIONING_README.md
- OmniVoice: https://github.com/k2-fsa/OmniVoice
- CodeSOTA TTS: https://www.codesota.com/text-to-speech
- Real speech sample: https://www.youtube.com/watch?v=EEFMVIf12UY