What TTS Throws Away: A Taxonomy and Exploratory Evaluation of Paralinguistic Script Controls in Modern Text-to-Speech Systems

Lissin Research

Draft preprint, June 2026

Abstract

Modern text-to-speech (TTS) systems can synthesize highly natural narration from plain text, but transcript and script-generation pipelines still discard many paralinguistic cues: held vowels, non-verbal reactions, filled pauses, mid-thought pivots, affect tags, pacing cues, and speaker-level delivery notes. We study whether restoring these cues in script form produces measurable naturalness gains when the TTS engine is already strong. We introduce a ten-axis taxonomy of script-side paralinguistic controls, map it across five content registers, and run an exploratory script-stripping ablation on ten production scripts. The ablation renders original and stripped variants through a Gemini-family TTS system and evaluates them with two Gemini-family audio judges. Under the ceiling-heavy Pro judge, original scripts average 4.65 MOS against 4.30 MOS for all-stripped scripts, a 0.35 MOS gap. As a descriptive five-point scale index, that moves from 86% to 93%; it is not a calibrated percentage of human naturalness. Under the Flash judge, the aggregate gap is much smaller. No per-axis confidence interval excludes zero. Transcript-density and engine-lane checks suggest that engine choice, voice choice, genre discipline, and evaluation design can move as much as or more than markup density. We therefore present the result as an exploratory taxonomy and evaluation, not as a definitive benchmark. The practical conclusion is that paralinguistic scripting remains a useful residual control layer, but external naturalness claims require blinded human-listener validation across engines and voices.

1 Introduction

Automatic speech recognition and transcript-oriented writing systems normalize speech into clean text. That cleaning is useful for readability, but it removes cues that matter to performance: stretched vowels, backchannels, laughter timing, sighs, hesitation, emphasis, pacing, affect shifts, and speaker routing. Companion qualitative examples are discussed in David [10]. The motivating question for this work is not whether those cues exist in human speech; prior work in discourse analysis, non-verbal vocalization, filled pauses, and expressive speech has already established that they do [1, 4, 8, 9, 13]. The narrower question is whether reintroducing those cues as script-side controls measurably improves synthesized speech when the TTS engine is already strong.

The first-order product intuition is simple: richer scripts give speech generation models more information for rendering a natural performance. The empirical difficulty is that modern engines already carry a large amount of naturalness on their own. A strong TTS model may infer prosody from ordinary punctuation, voice selection, and genre framing, while a weak or inconsistent control surface may ignore detailed markup. In other words, script markup is best treated as a residual-improvement layer on top of a high baseline, not

as the whole explanation for expressive speech. The result is a multi-factor problem: script quality, engine choice, voice choice, content register, prompt format, and evaluation design are entangled.

We make five contributions:

1. A ten-axis taxonomy for script-side paralinguistic controls relevant to generated narration.
2. A five-register usage matrix covering meditation, podcast, deep-dive/explainer, comedy-news, and lyrics.
3. An exploratory N=10 script-stripping ablation over original, all-stripped, and single-axis stripped variants.
4. A transcript-density audit of real human content that separates lexical signal from ASR-normalized orthographic artifacts.
5. A claim boundary that separates supported observations from hypotheses requiring human-listener validation.

The central finding is deliberately modest but still useful. Current TTS engines appear to provide most of the perceived naturalness in this setup; script markup adds a smaller, targeted residual gain. Under the Pro judge, the all-stripped condition still scored 4.30/5, while the original scored 4.65/5. Read only as a raw five-point score index, that is roughly 86% versus 93%, or a seven-point movement. This is not a literal share of human naturalness, but it gives the right product framing: do not exaggerate markup; use it to recover the last layer of timing, affect, emphasis, and register discipline. Per-axis attribution remains unstable, so the work supports a taxonomy, an evaluation harness, and a prioritization direction rather than a definitive external benchmark claim.

2 Background

2.1 Paralinguistic information in text

Written transcripts underrepresent many features that listeners use to infer stance, timing, and affect. Word lengthening in social text correlates with subjectivity and sentiment [5]. Filled pauses alter perceived fluency and processing [4, 8]. Discourse markers and conversational fillers organize turn-taking, stance, and topical shifts [1, 13]. Non-verbal vocalizations, including laughs, sighs, gasps, and other vocal bursts, carry affective information that is not reducible to words [9]. Expressive TTS therefore needs either to infer these cues from context or receive them explicitly in the input.

2.2 Modern TTS control surfaces

Recent speech models increasingly combine large language model structure with neural audio codecs and acoustic decoders [2, 3, 11, 18, 19]. Several recent research systems explicitly report human-parity or on-par quality on standard zero-shot TTS benchmarks [7, 15]. That literature supports the high-baseline framing: for frontier systems, the remaining question is often how to steer an already expressive model, not how to make an otherwise flat model expressive from scratch. Commercial TTS products expose different control interfaces: bracketed audio tags, SSML-like tags, natural-language style instructions, multi-speaker prompts, and voice catalogs [6, 12, 14, 17]. These controls are not equivalent. A bracket tag such as `[laughs]` is not the same interface as a natural-language instruction to speak playfully, and neither guarantees reliable rendering.

Table 1: Prior work and the gap addressed by this exploratory study.

Literature bucket	What it establishes	Gap addressed here
Discourse and disfluency	Filled pauses, discourse markers, and conversational fillers affect fluency, processing, turn structure, and stance [1, 4, 8, 13].	These cues are usually normalized away in clean scripts; we test what happens when script variants preserve or remove them.
Non-verbal vocalization	Laughs, sighs, gasps, and other vocal bursts carry affective information distinct from words [9].	TTS control surfaces expose these cues unevenly, and accepting a tag does not guarantee reliable rendering.
Modern TTS systems	Neural codec and language-model-based systems can already produce highly natural speech [2, 3, 7, 11, 15, 18, 19].	High baseline quality reduces the room for script markup to explain naturalness by itself.
Expressive TTS benchmarks	Newer evaluations test complex prosody and expressiveness across systems [16].	Broad expressiveness does not isolate the residual value of explicit script-side controls.
Vendor control surfaces	Commercial systems expose tags, SSML-like controls, style instructions, and voice settings [6, 12, 14, 17].	The interfaces are not equivalent; we audit them by axis rather than treating TTS systems as interchangeable renderers.

2.3 Evaluation risk

MOS-style evaluation is common in speech synthesis, but high-quality TTS can saturate five-point scales. Model-based audio judges add another risk: they may share family-level biases with the generator and may reward script features that resemble their own instruction-following conventions. We therefore use LLM audio judgments only as an exploratory signal. Human listener validation remains the threshold for publishable naturalness claims.

2.4 Prior work and the gap we address

Recent TTS evaluation work is moving beyond single-speaker naturalness and toward broader expressive capability. EmergentTTS-Eval, for example, constructs a large suite of samples for evaluating emergent expressive TTS behavior across tasks and systems [16]. That direction is important: if current TTS models can already generate plausible human-like prosody, evaluation pressure moves toward controllability, reliability, and use-case-specific expression rather than plain read-aloud quality alone.

The gap we address is narrower than broad expressiveness. Existing expressive TTS work asks whether a model can produce complex, expressive speech from a prompt. We ask how much explicit script-side markup changes the output when a modern engine already synthesizes a strong baseline from normalized text. This distinction matters for production systems. A model can score well on broad expressiveness while still failing to honor a specific pause, sigh, mid-thought pivot, or emphasis cue. Conversely, a model can produce pleasant speech while making detailed markup unnecessary.

Table 1 summarizes the literature buckets used to shape the study and the specific gap this work probes.

Table 2: Ten paralinguistic script-control axes used in the exploratory evaluation.

Axis	Script surface	Intended speech effect
Vowel elongation	`soooo`, `waaait`	Duration, emphasis, playful affect
Non-verbal vocalization	`[laughs]`, `[sighs]`, `[gasps]`	Audible reactions and affective vocal events
Filled pauses	`uh`, `um`, `hmm`	Hesitation, planning, conversational realism
Discourse markers	`so`, `well`, `yeah`, `right`	Turn framing, stance, topic flow
Conversational fillers	`you know`, `I mean`, `kind of`	Informal register and conversational texture
Mid-thought pivots	em dash, false starts, restarts	Self-correction and live reasoning
Emotion/delivery tags	`[curious]`, `[deadpan]`, `[whispers]`	Voice color and affective stance
Pacing tags	`[short pause]`, ellipses, slow/fast notes	Rhythm, silence, breathing room
In-text emphasis	capitalization, asterisks, italics	Local stress and focus
Audible reactions	`oh!`, `wow`, `ah`	Short exclamatory reaction tokens

3 Taxonomy

Table 2 gives the ten axes used throughout this work. The taxonomy is intentionally script-facing: it asks what a script writer can place in text before synthesis, not what a waveform analysis model might measure after synthesis.

The axes are not uniformly important across content registers. Comedy-news scripts often depend on timing, delivery tags, discourse markers, and pivots. Meditation scripts depend more on pacing and delivery, with less tolerance for filled pauses. Podcast scripts need a broader conversational surface. Figure 1 visualizes the register matrix used to guide the ablation and later product decisions.

4 Evaluation Design

4.1 Script-stripping ablation

The main exploratory experiment uses ten production scripts selected from a paralinguistically rich script set. For each script, we generate the original version, an all-stripped version, and ten single-axis stripped variants. Each stripped variant removes one surface class while leaving other script content intact. The full design therefore contains twelve variants per script.

The TTS rendering lane uses a Gemini-family TTS model with a fixed voice. Audio is then evaluated by two Gemini-family audio judges: a Pro judge and a Flash judge. Each judge scores four dimensions on a 1–5 scale: prosody naturalness, pacing naturalness, paralinguistic realism, and overall listenability. The reported MOS is the mean of these dimensions.

This design is intentionally described as exploratory. The scripts are not a preregistered random sample, the judges are not human listeners, and the generator and judges may share provider-level biases.

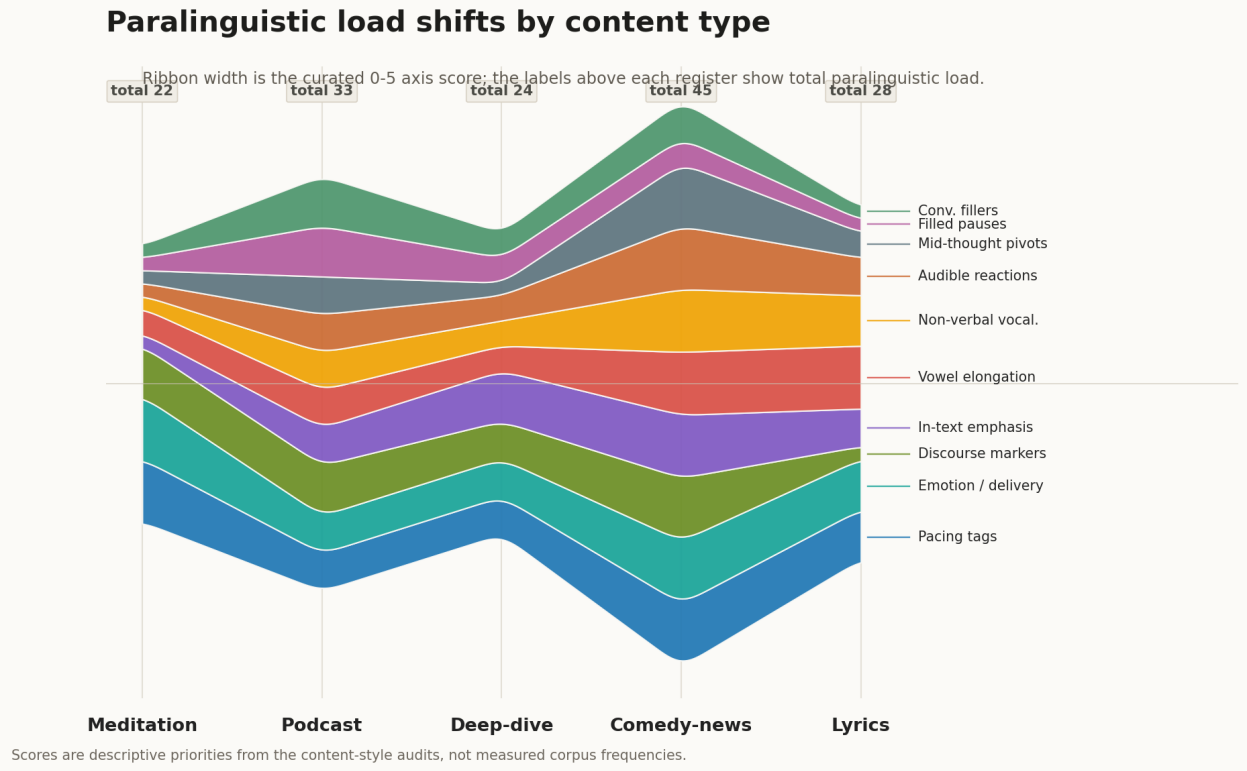

Figure 1: Paralinguistic axis usage by content register, scored from 0 (unused) to 5 (defining). The matrix is an editorial taxonomy used to guide evaluation design, not a measured corpus statistic.

4.2 Real-content transcript audit

To contextualize the script densities, we also audit 31 YouTube transcripts across meditation, podcast, comedy, and deep-dive content. The audit applies the same text-level detectors to ASR-derived transcripts. This comparison is useful only for axes that ASR reliably transcribes as ordinary words, such as discourse markers and conversational fillers. It is not valid for typography-dependent axes such as bracket tags, capitalization, asterisks, em dashes, and held-vowel spelling, because ASR normalizes or drops those surfaces.

4.3 Engine and pipeline checks

The report also includes non-confirmatory checks across engine lanes and production pipelines. These checks ask whether changing the TTS lane or production path can produce differences comparable to the script-stripping gap. They are not treated as a formal cross-engine benchmark because model versions, voice choices, and closed-system controls are moving targets.

5 Analysis Measures

The analysis uses three descriptive measures. First, we compute the mean MOS composite for each variant by averaging four model-judge dimensions: prosody naturalness, pacing naturalness, paralinguistic realism, and overall listenability. Second, we compute the aggregate original-minus-stripped gap for each judge. Third, we compute per-axis deltas by comparing each single-axis stripped variant with its original script.

Table 3: Headline exploratory ablation result. MOS is the four-dimension composite score on a 1–5 scale.

Quantity	Phase E, N=5	Corrected run, N=10
Mean MOS, original	4.75	4.65
Mean MOS, all-stripped	4.50	4.30
Original minus all-stripped	0.25	0.35
Cronbach's alpha	0.967	0.962

These measures are intentionally modest. They are useful for reading direction and scale, but they do not constitute a human MOS study. We therefore use them to identify where explicit script markup appears to help and where the evidence remains unstable.

6 Results

6.1 Aggregate script-stripping result

Table 3 reports the core ablation summary. The corrected N=10 run shows a 0.35 MOS gap between original and all-stripped scripts under the Pro judge. Cronbach's alpha over the four scoring dimensions is high, but high internal consistency among model-scored dimensions is not the same as human agreement.

The 0.35 MOS difference is small relative to the scale and should not be converted into a percentage of human naturalness. MOS has no meaningful zero point, and these scores come from model judges rather than listeners. Still, a descriptive scale-index reading is useful for product reasoning: all-stripped scripts scored 4.30/5, or 86% of the raw five-point scale, while original scripts scored 4.65/5, or 93%. The observed markup gain is therefore better described as a roughly five-to-seven point residual improvement on an already high baseline, not as the main source of naturalness.

6.2 Per-axis attribution is unstable

Figure 2 shows the per-axis deltas when each axis is stripped. Negative values indicate that stripping the axis reduced MOS relative to the original. Conversational fillers, emotion/delivery tags, and in-text emphasis have the largest point estimates, but no per-axis confidence interval excludes zero. These rankings are therefore hypotheses for a larger study, not confirmed causal effects.

6.3 Judge effects matter

The Pro judge gives ceiling-heavy scores, while the Flash judge produces a narrower and less favorable separation between original and all-stripped variants. Figure 3 compares the two judge families. The disagreement is a warning sign: relying on one model judge overstates the certainty of the result.

6.4 Engine and production changes can be larger

Later validation checks suggest that changing the TTS lane and production path can move observed quality as much as or more than removing markup. Figure 4 shows a five-way comparison that includes real human audio, alternate TTS lanes, a prompt-engineered script lane, and the deployed production pipeline.

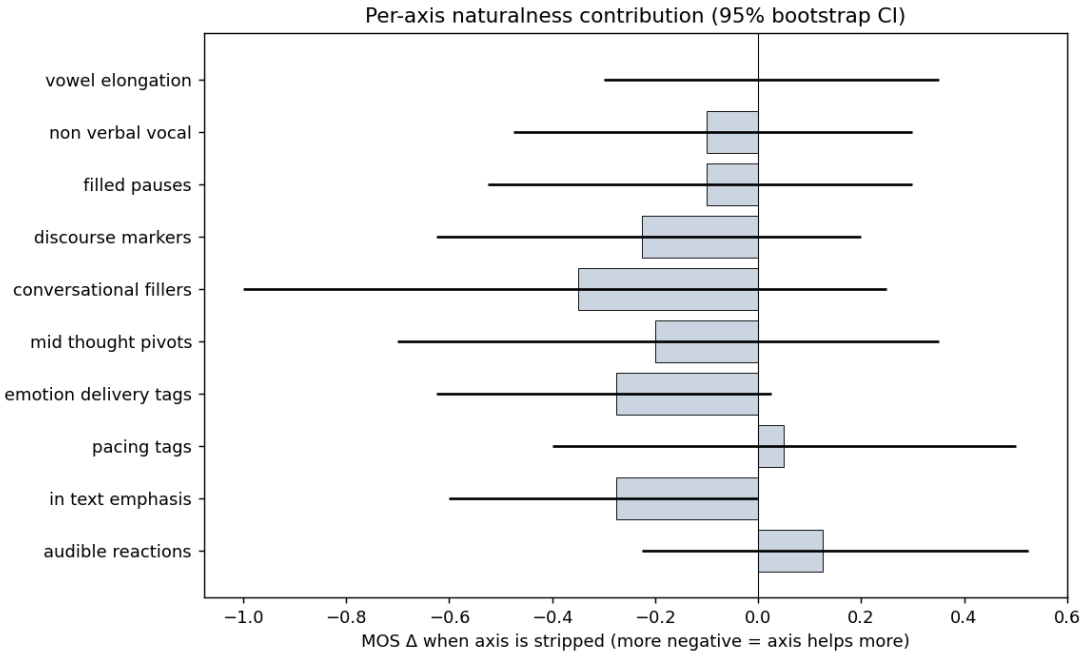

Figure 2: Per-axis MOS deltas from the N=10 exploratory ablation. No confidence interval excludes zero, so the axis ordering should be read as candidate prioritization rather than confirmation.

These comparisons are not controlled enough to be read as a benchmark, but they are important for product interpretation: script markup is one lever, not the only lever.

6.5 Real-content density audit

The real-content transcript audit shows that the lexical axes are closer to real human transcripts than the typographic axes suggest. Discourse markers and conversational fillers differ by roughly single-digit per-100-word amounts. In contrast, bracket tags, em dashes, all-caps emphasis, asterisks, and held-vowel spellings are not comparable against ASR transcripts because ASR normalizes them away. This supports a methodological distinction between lexical controls and script-format controls.

7 System Control Surfaces

We treat modern TTS systems as heterogeneous control surfaces rather than interchangeable speech renderers. Some systems expose documented bracket tags, some expose SSML-like controls, and some rely on natural-language instructions. Figure 5 summarizes the axis coverage pattern used in the system audit.

This distinction affects evaluation. A system that accepts `[laughs]` is not necessarily better than a system that infers a laugh from dialogue context; conversely, a system that accepts a tag may render it inconsistently. The audit therefore supports the same claim boundary as the ablation: control availability is not the same as control reliability, and naturalness is not explained by script markup alone.

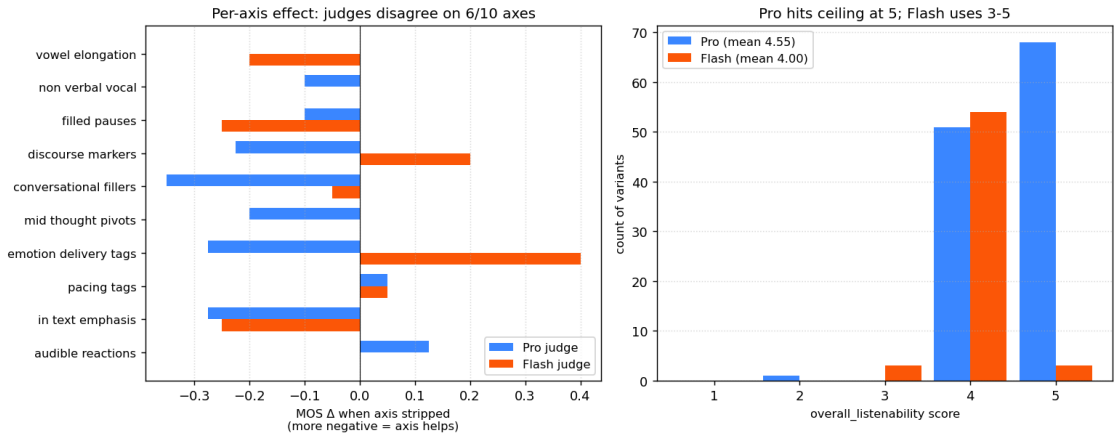

Figure 3: Model-judge comparison. The Pro judge gives more ceiling-saturated scores, while the Flash judge sees a much smaller original-vs-stripped gap.

8 Discussion

The main contribution is the claim boundary. The data-loss diagnosis is strong: clean transcripts and plain scripts discard paralinguistic information. The current exploratory evaluation does not show that markup is the dominant source of naturalness under modern TTS. Instead, it supports a residual-improvement view: current engines already produce high-scoring speech from stripped scripts, while script markup contributes targeted gains in the last layer of timing, affect, emphasis, and genre fit. Per-axis attribution is not reliable at N=10, and engine and evaluation design are large enough to change interpretation.

This is still useful. A script writer should preserve paralinguistic cues because they provide an explicit control interface, especially for genres where timing, affect, and reactions are defining. But the product decision is not to maximize markup density. The more defensible product interpretation is to choose the right engine and voice, keep genre-specific markup disciplined, and evaluate the final audio with listeners.

Table 4 summarizes which claims are currently supported and which require further validation.

9 Limitations

The study has several important limitations.

- **Model judges are not listeners.** The evaluation uses Gemini-family audio judges, not blinded humans.
- **Same-family bias is possible.** Generator and judges may share provider-level preferences.
- **MOS is ceiling-clipped.** Strong TTS outputs can saturate a five-point scale.
- **N=10 is exploratory.** Per-axis effects are underpowered and should be treated as hypotheses.
- **Scripts are not randomly sampled.** The scripts come from a paralinguistically rich production set.
- **Closed TTS systems drift.** Vendor model versions, voices, and undocumented control behavior can change.

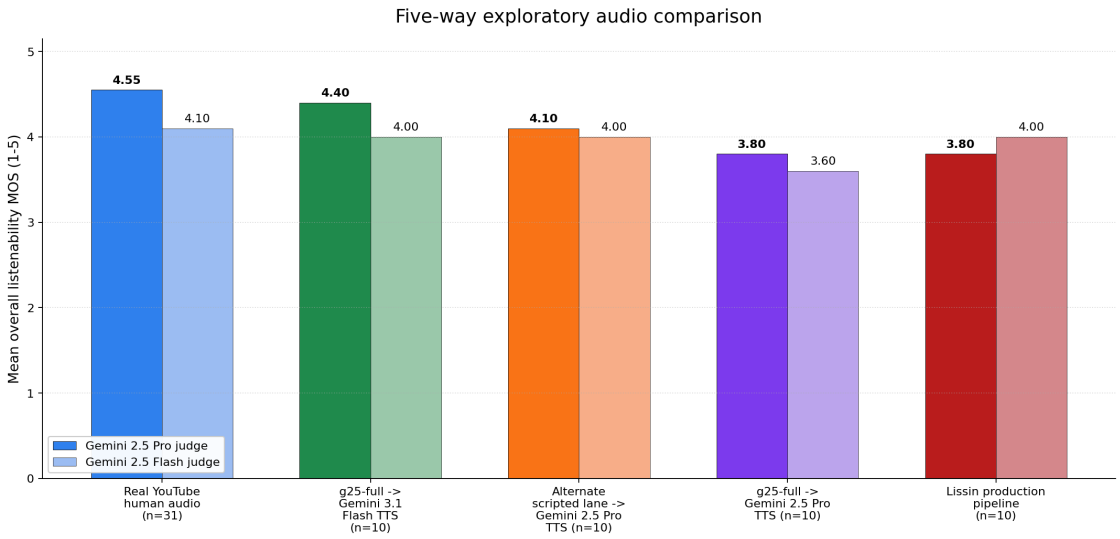

Figure 4: Five-way audio comparison from the validation phase. The result motivates joint evaluation of script, voice, engine, and production layer rather than script-only claims.

- **Transcript-density comparisons are axis-dependent.** ASR transcripts normalize typography and should not be used as ground truth for bracketed or orthographic controls.
- **The ten-axis taxonomy is incomplete.** It is intentionally script-facing and intentionally single-speaker, single-track. Several real phenomena that occur in podcast, comedy, and storytelling audio fall outside the ten axes. Section 10 lists the most important omissions identified during validation.

10 Taxonomy Boundary

The ten axes in Table 2 were derived from a specific subset of paralinguistic phenomena: speaker-internal markers that appear at the word or phrase level in a single-speaker single-track recording. Several phenomena in real podcast, comedy, and storytelling audio are not captured by any of the ten. Table 5 lists the most important omissions identified during validation. We keep them separate from the core taxonomy because the current ablation did not test them.

This boundary matters because several audio formats that sound natural to listeners are not well described by single-speaker script markup alone. The taxonomy covers a useful layer of the problem, not the full production stack.

11 Reproducibility

The current project contains the script variants, rendered audio, model-judge outputs, figure-generation scripts, and paper source used for this preprint. The appendix records the transformation classes and the prompt text used for the model-judge and A/B validation lanes. Reproduction has three levels:

The main reproducibility limit is model drift. Closed TTS systems and model judges can change over time, and voice snapshots may not be externally frozen. For that reason, we treat the present numbers as exploratory evidence from a dated run rather than as a stable public score.

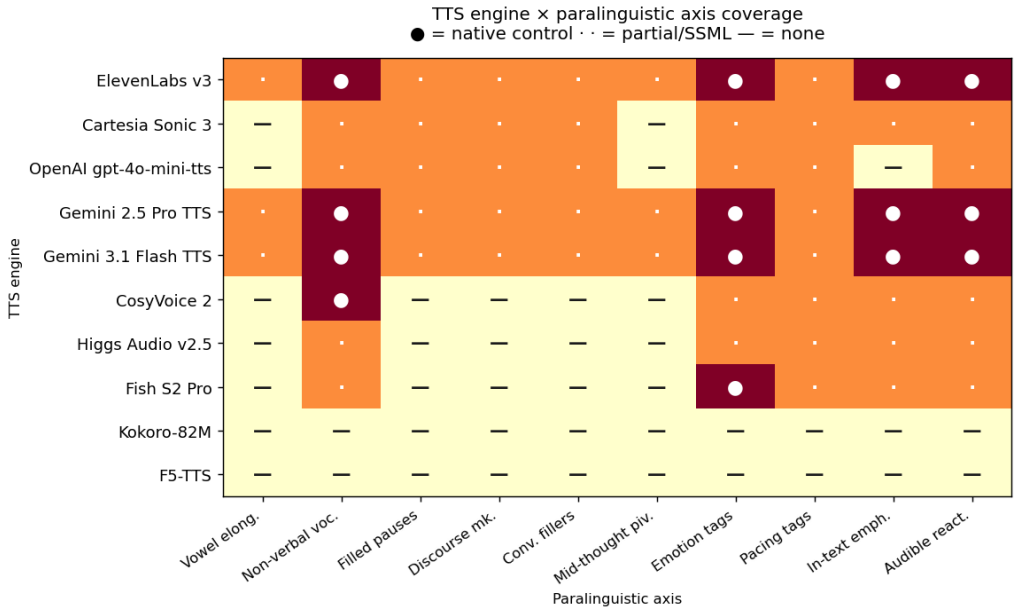

Figure 5: Illustrative axis coverage across representative TTS control surfaces. Scores compress very different interfaces and do not measure tag reliability.

12 Conclusion

Paralinguistic scripting is a real and useful control surface for modern TTS, but the current evidence supports a restrained conclusion. The ten-axis taxonomy helps structure script design and evaluation. The exploratory ablation shows small model-judge gains from preserving markup, but not stable per-axis causal effects. Engine choice, voice choice, production layer, and judge design remain first-order variables. A defensible external naturalness claim has to evaluate scripts, engines, voices, and listeners together.

References

[1] Galina B. Bolden. Implementing incipient actions: The discourse marker so in english conversation. *Journal of Pragmatics*, 41(5):974–998, 2009. doi: 10.1016/j.pragma.2008.10.004.

[2] Zalán Borsos et al. Audiolm: A language modeling approach to audio generation, 2022. URL https://arxiv.org/abs/2209.03143.

[3] Zalán Borsos et al. Soundstorm: Efficient parallel audio generation, 2023. URL https://arxiv.org/abs/2305.09636.

[4] Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147, 2001. doi: 10.1177/00238309010440020101.

[5] Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, 2011. URL https://aclanthology.org/D11-1052/.

Table 4: Claim status after the exploratory evaluation.

Claim	Status	Reason
ASR and clean transcripts discard paralinguistic cues	Supported qualitatively	Consistent with literature and transcript inspection
Ten-axis taxonomy is useful for script design	Supported as design framework	Axes map to observable writing and TTS controls
Script stripping produces a measurable aggregate model-judge gap	Exploratory support	0.35 MOS under one judge, much smaller under another
Script markup acts as a residual improvement layer	Supported as framing	All-stripped scripts remain high-scoring; markup shifts the raw scale index by roughly five-to-seven points
Single axes have confirmed causal effects	Not supported yet	No per-axis confidence interval excludes zero
Script markup dominates engine choice	Not supported yet	Engine and production-lane changes can be comparable or larger
External naturalness claim	Not ready	Requires blinded human-listener study

[6] Cartesia. Sonic text-to-speech model documentation. https://docs.cartesia.ai/build-with-cartesia/tts-models/latest, 2026. Accessed June 2026.

[7] Sanyuan Chen et al. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers, 2024. URL https://arxiv.org/abs/2406.05370.

[8] Martin Corley and Oliver W. Stewart. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602, 2008. doi: 10.1111/j.1749-818X.2008.00068.x.

[9] Alan S. Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698–712, 2019. doi: 10.1037/amp0000399.

[10] Amal David. What tts throws away. https://amaldavid.com/writing/what-tts-throws-away/, 2026. Companion essay.

[11] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022. URL https://arxiv.org/abs/2210.13438.

[12] ElevenLabs. Eleven v3 audio tags. https://elevenlabs.io/blog/v3-audiotags, 2025. Accessed June 2026.

[13] Bruce Fraser. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952, 1999. doi: 10.1016/S0378-2166(98)00101-5.

[14] Google AI for Developers. Gemini api speech generation documentation. https://ai.google.dev/gemini-api/docs/speech-generation, 2026. Accessed June 2026.

[15] Zeqian Ju et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models, 2024. URL https://arxiv.org/abs/2403.03100.

Table 5: Paralinguistic phenomena outside the ten-axis taxonomy used in this study.

Phenomenon	Why it is outside the current taxonomy
Onomatopoeia	Comedy and storytelling often use lexical imitation of external sounds (`BOOM`, `tap-tap-tap`); the ten axes focus on speaker-internal cues.
Overlapping speech and turn-taking	Conversational podcasts include simultaneous speech and clipped turns; the current study uses single-track script variants.
Persona switching	Stand-up, narration, and storytelling often switch character voice mid-utterance; this requires speaker-state control beyond local tags.
Backchannels during another turn	Interview audio includes listener backchannels on a parallel conversational track; the audible-reactions axis only covers speaker-emitted tokens.
Sound effects and foley cues	Audio dramas and story podcasts write cues such as `[door slam]` or `[footsteps]`, but those are separate sound layers rather than TTS voice events.
Dynamic range within an utterance	Whisper-to-shout motion is a continuous acoustic control, not just a discrete delivery tag.
Code-switching and accent shift	Multilingual delivery and accent shifts are scriptable, but the current English single-voice setup does not test them.
Audience reaction integration	Comedy audio often depends on a third-party audience track, which is outside single-speaker TTS rendering.
Sub-segmental phonation	Vocal fry, creaky voice, breathiness, and reduction patterns live below the granularity of the script-facing axes used here.

Table 6: Reproducibility layers for the current exploratory study.

Layer	Local artifacts	Purpose
Text variants	Axis-stripping code and variant text files	Reconstruct original, all-stripped, and single-axis stripped scripts
Audio renders	Rendered WAV/MP3 files and lane labels	Inspect the clips used by the audio judges
Model judgments	JSON score files and aggregation tables	Recompute MOS summaries, per-axis deltas, and judge comparisons
Paper build	LaTeX source, figures, bibliography, and Makefile	Rebuild the PDF and inspect the written claim boundary

[16] Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge, 2025. URL https://arxiv.org/abs/2505.23009.

[17] OpenAI. Text to speech guide. https://developers.openai.com/api/docs/guides/text-to-speech, 2026. Accessed June 2026.

[18] Paul K. Rubenstein et al. Audiopalm: A large language model that can speak and listen, 2023. URL https://arxiv.org/abs/2306.12925.

[19] Chengyi Wang et al. Neural codec language models are zero-shot text to speech synthesizers, 2023. URL https://arxiv.org/abs/2301.02111.

A Content-Register Matrix

Table 7 gives the same editorial matrix as Figure 1. Scores range from 0 (unused) to 5 (defining). The matrix is a design prior used to select and interpret axes; it is not estimated from a corpus.

Table 7: Axis usage by content register.

Axis	Meditation	Podcast	Deep-dive	Comedy-news	Lyrics
Vowel elongation	2	3	2	5	5
Non-verbal vocalization	1	3	2	5	4
Filled pauses	1	4	2	2	1
Discourse markers	4	4	3	5	1
Conversational fillers	1	4	2	3	1
Mid-thought pivots	1	3	1	5	2
Emotion/delivery	5	3	3	5	4
Pacing tags	5	3	3	5	4
In-text emphasis	1	3	4	5	3
Audible reactions	1	3	2	5	3

B Ablation Variants

Each script was rendered as one original version, one all-stripped version, and ten single-axis stripped variants. The axis-specific transformations were text-level removals or normalizations. Table 8 summarizes the intended transformation class.

Table 8: Script transformation classes used by the ablation.

Variant	Transformation intent
All stripped	Remove all ten paralinguistic surfaces while preserving core lexical content
Vowel elongation stripped	Normalize repeated-character vowel spellings to ordinary spellings
Non-verbal vocalization stripped	Remove bracketed vocal events such as laughs, sighs, gasps, and coughs
Filled pauses stripped	Remove lexical pauses such as uh, um, and hmm
Discourse markers stripped	Remove selected discourse framing tokens when not semantically essential
Conversational fillers stripped	Remove filler phrases such as you know, I mean, and kind of
Mid-thought pivots stripped	Normalize false starts, restarts, and dash-marked pivots
Emotion/delivery stripped	Remove bracketed delivery and affect tags
Pacing stripped	Remove explicit pause and pacing annotations
In-text emphasis stripped	Normalize capitalization, asterisk emphasis, and related local stress marks
Audible reactions stripped	Remove short exclamatory reaction tokens where possible

C Per-Axis Exploratory Results

Table 9 gives the corrected N=10 per-axis deltas. Delta is defined as stripped variant minus original; lower values indicate that the stripped version scored worse. None of these intervals excludes zero.

Table 9: Per-axis exploratory deltas, N=10.

Axis	N	Mean delta	95% CI	p
Vowel elongation	10	+0.00	[-0.30, +0.35]	1.000
Non-verbal vocalization	10	-0.10	[-0.47, +0.30]	0.641
Filled pauses	10	-0.10	[-0.53, +0.30]	0.641
Discourse markers	10	-0.23	[-0.62, +0.20]	0.289
Conversational fillers	10	-0.35	[-1.00, +0.25]	0.272
Mid-thought pivots	10	-0.20	[-0.70, +0.35]	0.464
Emotion/delivery tags	10	-0.28	[-0.62, +0.03]	0.128
Pacing tags	10	+0.05	[-0.40, +0.50]	0.834
In-text emphasis	10	-0.28	[-0.60, +0.00]	0.086
Audible reactions	10	+0.12	[-0.23, +0.53]	0.555

D Control-Surface Interpretation

The system-surface audit reduces each TTS interface to three coarse categories:

- **None:** no documented or observed direct control for the axis.
- **Partial:** lexical, punctuation, style-instruction, or best-effort support.
- **Native:** documented tag, SSML-like control, or explicit audio-cue support.

This compression is useful for comparison, but it hides reliability. A native tag that renders inconsistently may be less useful than a partial natural-language instruction that renders reliably. The audit therefore separates documented availability from observed reliability wherever the evidence allows.

E Operational Prompts and Protocol Text

This appendix records the prompt and protocol text used by the experimental lanes that inform the preprint. It is included so the reader can inspect the actual evaluation setup rather than infer it from prose.

E.1 Script authoring prompt excerpt

The paralinguistically rich script lane used a system prompt that forced both bracketed tags and spoken disfluencies. The excerpt below shows the constraint style.

```
You write spoken-podcast scripts for Google Gemini 2.5 Pro TTS. The script you
produce MUST contain BOTH of the following - they are equally required, not
alternatives.

REQUIREMENT 1 - Inline bracketed tags as prosodic directives.

Use ONLY tags from this Gemini 2.5 Pro TTS vocabulary:
- Emotion / delivery: [amazed], [angry], [bored], [crying], [curious],
  [empathetic], [excited], [excitedly], [furious], [mischievously],
  [panicked], [reluctantly], [sarcastic], [scornful], [serious], [tired],
  [trembling], [robotic], [whispering], [whispers], [speaking slowly],
  [shouting]
- Non-verbal vocal sounds: [laughs], [laughing], [giggles], [sighs],
```

```
  [sighing], [gasp], [clears throat], [cough]
- Pacing / pauses: [short pause], [medium pause], [long pause],
  [PAUSE=2s], [extremely fast], [very fast], [very slow]

Rules:
- Place one delivery/emotion tag near the start.
- Use 2-3 pacing/pause tags at phrase boundaries.
- Use 1-2 non-verbal vocalizations where natural.
- Aim for 5-10 tags in a 3-5 minute script.
- Tags go OUTSIDE words, before or around the phrase they affect.

REQUIREMENT 2 - Real podcast verbal disfluencies woven into the prose.

- 'uh', 'um', 'you know', 'I mean', 'like', 'sort of', 'I guess',
  'kind of' scattered through sentences.
- False starts and mid-thought pivots: "and so - and so the reason is...",
  "I was - well, what happened was..."
- Aim for 8-15 disfluencies in a 3-5 minute script.
- Polished broadcast English is unacceptable.

Output ONLY the script - no preamble, no explanation, no markdown.
```

E.2 TTS rendering protocol

Each original, all-stripped, and single-axis stripped variant was sent as raw script text to Gemini 2.5 Pro TTS with the fixed voice `Kore`. Renders were written as 24 kHz mono WAV files. For long scripts, the render call used the first 3,500 characters of the variant text so that every condition used the same truncation rule.

E.3 Audio MOS judge prompt

The following prompt was used for the main script-stripping MOS-style judgments.

```
You are evaluating a podcast/audio TTS sample for **naturalness** - how
human-sounding it is.

Rate the audio on each of these 4 dimensions, 1-5
(1=very robotic, 5=indistinguishable from a real podcast host):

- prosody_naturalness: pitch, rhythm, intonation contour
- pacing_naturalness: pause placement, tempo variation
- paralinguistic_realism: do disfluencies/breaths/laughs/tags sound real or
  pasted-on?
- overall_listenability: would you keep listening to this?

Respond as a single JSON object with the 4 scores and a one-sentence
"rationale" - nothing else.

Example: {"prosody_naturalness": 4, "pacing_naturalness": 3,
"paralinguistic_realism": 3, "overall_listenability": 4,
"rationale": "Pauses feel slightly long but pitch contour is convincing."}
```

E.4 Production-versus-prompt A/B judge prompt

The validation phase also used a pairwise A/B prompt to compare production audio with the prompt-engineered script lane.

```
You will hear two TTS audio samples (Audio A and Audio B). They were
synthesized from DIFFERENT scripts of the SAME source content - one written by
```

```
Lissin's existing production pipeline, one written by a prompt-engineered
frontier LLM. You don't know which is which.

Listen to both. Compare on:
1. paralinguistic_realism - do disfluencies / pauses / laughs sound real?
2. prosody_naturalness - pitch / rhythm / intonation
3. overall_listenability - would you keep listening?

For each, pick A, B, or TIE.

Respond as JSON: {"paralinguistic_realism": "A|B|TIE",
"prosody_naturalness": "A|B|TIE",
"overall_listenability": "A|B|TIE",
"rationale": "one sentence"}.

A is the FIRST audio, B is the SECOND.
```

F Example Variant Family

Table 10 shows the structure of one short variant family in the style used by the ablation.

Table 10: Illustrative prompt family.

Variant	Script text
Semantic core	The host realizes the guest has changed their mind and invites them to explain the shift.
Canonical	So, you were certain about this yesterday. What changed?
Marked	So, you were certain about this yesterday – wait, no, actually – what changed? `[curious, softer]`
All stripped	You were certain about this yesterday. What changed?
Target axis	Mid-thought pivot plus delivery tag
Expected cue	A short self-correction followed by softer, curious delivery

This item is not scored merely on whether the marked clip is more dramatic. The relevant questions are whether the false start is audible, whether the softened delivery fits the turn, and whether the result remains natural. If the system overacts the hesitation, the cue is realized but inappropriate.

G Failure Modes

The validation process identified the following common failure modes for script markup:

- **Cue omission:** the model ignores the target cue entirely.
- **Cue literalism:** the model reads a bracketed cue aloud or treats markup as ordinary text.
- **Overacting:** the cue is audible but too theatrical for the register.
- **Semantic drift:** the model changes the words or intent while trying to satisfy the style cue.
- **Timing drift:** the cue appears in the wrong location.
- **Voice instability:** the cue changes speaker identity, pitch character, or accent unexpectedly.

- **Retry dependence:** the cue appears only after repeated generation attempts.

These labels were not scored as a formal metric in the N=10 ablation. They are included to make the qualitative reading of failures more concrete.