

Can Prosodic Style Be Inferred from Text Alone? Evidence from Unsupervised Acoustic Clusters

Abdul Rehman, Jian-Jun Zhang, and Xiaosong Yang

Abstract—Much of expressive text-to-speech research rests on an untested assumption that written text carries enough information to select an appropriate prosodic style for its delivery. Text-predicted style models improve listener preference, and expressive-appropriateness evaluation presupposes that context constrains style, yet neither measures the assumption itself. This paper tests it as a falsifiable hypothesis against style labels derived from acoustics alone. For each of six speakers in a 1,200-hour conversational corpus, utterances are clustered in the spaces of five speech models, including a prosody-only control, and the cluster of held-out utterances is predicted from twelve text embedding models. Three controls are applied: utterance length is erased from the speech embeddings; accuracy is scored against the majority-class floor of unbalanced clusters rather than uniform chance; and a bag-of-words baseline measures word identity alone. Text predicts the cluster above that floor for all six speakers (+0.111 top-3 accuracy), but bag-of-words achieves three quarters of this. Sentence embeddings add only +0.026, largest for encoders *not* trained for sentence semantics and reversed by tree-based probes for all others. Acoustic clusters are not compact in text embedding space in any of 360 configurations. The prosody-only space weakens the association for five speakers, but not for the speaker showing it most strongly. Text thus informs these delivery clusters mainly through word choice, whether as a cue to prosody or as a marker of topic and recording situation, and reference-free style selection cannot assume more.

Index Terms—Prosody, speech synthesis, sentence embeddings, style selection, text-to-speech.

I. INTRODUCTION

HUMAN speech communicates through two coupled channels: the linguistic content (*what* is said) and its paralinguistic realization (*how* it is said). Modern text-to-speech (TTS) systems handle the first channel well but rely on external conditioning for the second: to obtain contextually appropriate prosody, many systems require reference audio, explicit style labels, or manually selected style tokens [1]–[5]. A system asked to read an apology, a sponsor announcement, and a casual aside should be able to infer from the text that each calls for a different delivery, yet without explicit style conditioning it typically cannot.

A substantial body of work suggests that this bottleneck is solvable from the text side. A style representation predicted from text improves listener preference over an unconditioned baseline [6], and a growing family of successors predicts style or prosody from progressively richer text-side inputs: cross-sentence context [7], [8], hierarchical emotion distributions deduced solely from text [9], contrastively learned text–prosody alignments [10], and diffusion models that sample prosody

conditioned on text [11], [12]. Work on expressive appropriateness likewise shows that human judgments of whether a rendering is suitable depend on discourse context rather than on the isolated waveform [13]. All of these results, however, presuppose rather than measure the underlying premise: the prediction targets are latent style embeddings or prosodic features co-trained with the synthesis model itself, and the outcome measures are synthesis quality and listener preference. None of this work quantifies how strongly text constrains prosodic style against an explicit null, separates a text–prosody association from the two trivial explanations that any such measurement admits, utterance length and unbalanced label priors, or measures how much of it is carried by word identity alone. More broadly, much of TTS research implicitly expects style selection from text to emerge given enough model scale and data [14], [15]; that expectation has, to our knowledge, never been tested directly. This paper provides that test.

The question addressed is: **can prosodic style be inferred from text alone?** Concretely, if the prosodic styles a speaker actually uses are identified from acoustics without any text supervision, can those styles be predicted from a text-only representation of what was said? The defensible form of this hypothesis is probabilistic. A deterministic mapping from text to prosody is arguably impossible: the same sentence admits multiple contextually valid renderings depending on discourse context and speaker intent, which are not recoverable from the text itself [16]. What is tractable is the weaker claim that textual content is not prosodically *neutral*, consistent with long-standing evidence that intonation and speaking style reflect discourse structure and communicative intent [17], [18]. The hypothesis under test and its null are:

H_1 : Within an individual speaker’s natural speech, textual content statistically constrains prosodic style: a text-only representation of an utterance predicts its acoustically derived prosodic style category better than utterance length and label frequency do, and those categories occupy distinguishable regions of text embedding space.

H_0 : Prosodic style, as identified by unsupervised clustering of acoustics, is independent of text beyond utterance length and label frequency: text representations do not beat those references, and the style categories are geometrically indistinguishable in text embedding space.

Word identity is treated as a reference level rather than as a confound. Individual words can themselves cue delivery, as interrogatives, intensifiers, and discourse markers do, so an association carried by word identity would still be usable by a TTS system; but the same association would also arise if a speaker’s vocabulary merely tracks topic and recording

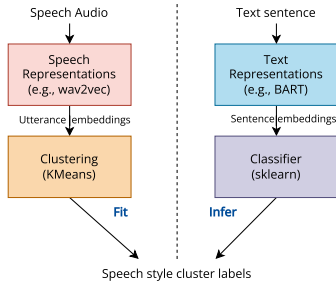


Fig. 1. The cross-modal evaluation framework. Utterances are clustered using speech representation models (Table I); the cluster labels are then predicted from text embeddings of the transcriptions, and compared against length-only and bag-of-words arms that see the same labels and the same held-out utterances.

situation. A bag-of-words arm is therefore reported alongside the sentence embeddings, and the question whether text is informative at all is kept separate from the question whether sentence-level representations add anything beyond the words they contain.

The hypothesis is operationalized as a cross-modal cluster prediction task (Fig. 1). Each speaker’s utterances are embedded by a speech representation model and clustered; the cluster index serves as an acoustically derived prosodic style label. Utterance embeddings are mean-pooled over segments of up to 20 s, so these labels describe utterance-level delivery, the register in which a stretch of speech is produced, which is also the level at which reference encoders and style tokens condition a TTS system [3]; local intonation contours such as a final question rise are outside the scope of the design. The transcriptions of the same utterances are embedded by a sentence transformer, and H_1 is confronted with three falsifiable predictions: (P1) text predicts the acoustic cluster of held-out utterances above the majority-class floor and the length-only baseline, with sentence embeddings adding to what bag-of-words achieves; (P2) the acoustic clusters occupy distinguishable, even if overlapping, regions of the text embedding space; and (P3) the alignment does not arise from lexical content leaking through the speech representation, and therefore persists when that representation contains no lexical channel. Each prediction can fail independently. P3 addresses the most serious confound of the design: general-purpose speech representations encode phonetic, and hence lexical, content, so an apparent text–speech alignment can be lexical leakage rather than prosodic coupling. The speech-model pool therefore includes, alongside four self-supervised models, a deliberately minimal prosody-only model (TIMBRESpace, Section III) that discards phonetic detail by construction.

The main contributions are:

- 1) To our knowledge, the first test of the hypothesis that prosodic style is inferable from text against independently derived (unsupervised, acoustics-only) style labels rather than co-trained latents. The assumption exploited, implicitly or by construction, in prior work [6]–[13] is stated with an explicit null and evaluated across 5 speech representation models, 12 text embedding models, 6 classifier families, and 3 encoder depths,

under a strict within-speaker protocol on six speakers. The speech encoders behind the labels are separately validated on labeled speech (Section VII), where the same frozen representations decode four-way emotion at 0.53 to 0.59 weighted F_1 against a 0.197 majority floor.

- 2) A measurement protocol with the three controls such a test requires, and evidence that they are not optional: length-conditional erasure of the speech embeddings before clustering, which brings the length-only baseline down to the level of a constant predictor; scoring against the majority-class floor, which absorbs more than half of the apparent margin over uniform chance; and a bag-of-words baseline, which absorbs three quarters of what remains.
- 3) The resulting bounded result. P1 survives in a weak form: text predicts its own utterance’s acoustic cluster above the floor for all six speakers (+0.111 top-3), mostly through word identity, and sentence embeddings add only +0.026 over bag-of-words, positive for every speaker on average. P2 fails: acoustic clusters are not compact in text embedding space in any of 360 configurations.
- 4) Two diagnostics on the character of the association. On the text side, `roberta-base` and `distilbert-base-uncased`, which are not trained for sentence similarity, give the largest increments and are the only encoders whose advantage over bag-of-words survives tree-based probes, while the 0.6B–1.5B instruction-tuned embedders rank last, which argues against a semantic reading of the increment. On the speech side, P3 holds only in part: the prosody-only representation retains less than half of the waveform encoders’ association for five of six speakers, consistent with lexical leakage, yet most of it for the speaker with the strongest association, which lexical leakage cannot explain.

II. BACKGROUND

A. Self-Supervised Speech Representations

Self-supervised (SSL) speech models learn representations from raw waveforms that transfer well to supervised downstream tasks [19]. The three dominant families differ in their pretraining targets: `wav2vec 2.0` uses contrastive targets over masked frames [20], `HuBERT` predicts iteratively refined K-means cluster assignments [21], and `WavLM` adds speech denoising and augmentation to masked prediction [22]. The ProsAudit benchmark [23] shows that SSL models encode structural prosodic knowledge above chance, which supports treating their embedding spaces as a source of prosodic structure. Downstream benchmark success, however, does not imply that these models organize speech into coherent *unsupervised* structure: benchmark rankings are sensitive to probing-architecture choices [24], speaker and content information occupy largely orthogonal subspaces [25], [26], and an analysis of ten pretrained SSL models by Rehman *et al.* [27] found that strong classification performance frequently coexists with weak clustering coherence for the same attribute.

The present study responds with a within-speaker clustering protocol, which controls for the dominant speaker-identity structure in SSL embedding spaces, and with a prosody-specific control representation alongside the general-purpose waveform encoders.

B. Prosody and Style in TTS

Expressive TTS systems condition synthesis on style information obtained from reference audio encodings [1]–[3], explicit pitch and duration targets [28], [29], adversarial style discriminators built on SSL representations [4], large-scale speech-token language modeling [5], prosody-focused self-supervised objectives [30], [31], or discrete latent style spaces [32], [33]. All of these mechanisms answer *how* to impose a chosen style; none answers *which* style should be chosen for a given input text.

A second line of work predicts the style from the text itself. Text-predicted style models [6] train a predictor from text to the latent style embedding of a GST-Tacotron; hierarchical context-aware predictors extend the text-side input to neighboring sentences and previously realized styles [7]; emotion-distribution predictors deduce multi-level emotion intensities solely from text [9]; and diffusion-based predictors replace deterministic regression with sampling, motivated by the one-to-many character of the text-to-prosody mapping [11], [12]. Closest of all are contrastive frameworks that align text with prosody or style directly: CLAPSpeech [10] contrasts the prosody of the same spoken token across different text contexts, and TACA [8] learns an utterance-level text-aware style space under the supervision of a speech-derived one. Related threads predict style embeddings from natural-language style prompts [34], which supply the style description explicitly, and model prosody from joint acoustic and textual input, where the decoder is observed to rely chiefly on the textual stream [35]. The consistent success of this family establishes that *some* usable signal exists in text, but the evidence is indirect: the prediction targets are latent variables co-trained with the synthesis model or the alignment objective, the outcome measure is synthesis quality or subjective preference, and, with the partial exception of CLAPSpeech’s same-token contrast, nothing in these designs excludes lexical, length, or channel information as the carrier of the learned association. From the evaluation side, Wang *et al.* [13] formalize *expressive appropriateness*, whether given speech fits its discourse context, and show that context-free judgments consistently underperform context-aware ones; there, the text-to-style link is the premise of the framework, not its object of measurement. None of this work measures whether, and how strongly, text suffices to *choose* the style, which is the hypothesis test carried out here.

C. Sentence Embeddings

Sentence transformers [36] map text to dense vectors whose distances reflect semantic similarity, with variants trained on natural language inference [37], paraphrase pairs, and general-purpose contrastive objectives [38]–[40]. Large language models and retrieval-augmented generation [41], [42] would likely host an eventual deployment of text-driven style selection.

This study deliberately uses frozen text embeddings with shallow classifiers instead, isolating the question of whether the information is present and accessible from the engineering question of how best to exploit it, and it includes plain masked-language-model encoders alongside similarity-trained ones so that “semantic” and “surface” explanations of any effect can be told apart.

III. A MINIMAL PROSODY-ONLY REPRESENTATION

Testing whether text–speech alignment is carried by prosody requires a speech representation with “just enough” information about prosody and timbre and as little as possible about lexical content. None of the SSL baselines satisfies this: their waveform input retains full phonetic detail. TIMBRESpace is a simple transformer model built for this purpose; code and checkpoints are publicly available¹, and further implementation details are given in the supplementary material.

TIMBRESpace operates on six frame-level acoustic features extracted at 62.5 frames per second (50 ms analysis frames, 16 ms hop): fundamental frequency (F0), F0 bandwidth, first and second formant frequencies (F1, F2), RMS power, and a formant-to-remainder signal-to-noise ratio (SNR), each normalized to [0,1] over fixed, perceptually motivated ranges. The set covers the primary acoustic correlates of prosody and voice quality while excluding phoneme-discriminative representations such as MFCCs. F1 and F2 carry some vowel information, but the feature set as a whole is drastically less lexically informative than spectrogram or waveform input, which is what makes the model useful as a lexical-leakage control. The frames are quantized into a vocabulary of 120 prosodic tokens by weighted K-means, and an 8-layer, 8-head, 768-dimensional transformer encoder is trained with BERT-style masked language modeling over the token sequences, using span masking (1 to 10 tokens) and a curriculum masking probability increasing from 1% to 20% [21], [44]. Training uses LibriSpeech [45] in two phases (train-clean-360, then train-other-500). Utterance-level embeddings are obtained by mean pooling frame representations from a chosen layer, identically to the SSL baselines.

Two properties matter here, both measured directly in Section VII. First, TIMBRESpace embeddings cluster along paralinguistic dimensions despite, or because of, their restricted input. Second, speaker identity is the dominant axis in the space, as it is in SSL embedding spaces [27] and in TTS style-embedding spaces, which cluster first by speaker timbre and only then by style attributes [34]. Any cross-speaker clustering would therefore produce speaker partitions rather than style partitions, which motivates the strict within-speaker protocol used throughout this paper.

IV. METHODOLOGY

A. Task Definition

For each speaker, let $\{(a_i, x_i)\}_{i=1}^N$ denote utterances with audio a_i and transcription x_i . A speech representation model S produces utterance embeddings $S(a_i)$ (mean-pooled over

¹<https://anonymous.4open.science/r/Timbrespace/>

TABLE I
SPEECH REPRESENTATION MODELS

Model	Params	Input	Pretraining data	Objective
HuBERT Large [21]	300M	waveform, 16 kHz	60k h Libri-Light	masked prediction of K-means targets
HuBERT XLarge [21]	1B	waveform, 16 kHz	60k h Libri-Light	masked prediction of K-means targets
WavLM Large [22]	300M	waveform, 16 kHz	94k h mixed	masked prediction + denoising
wav2vec2 XLSR-53 [43]	300M	waveform, 16 kHz	56k h, 53 languages	contrastive
TIMBRESpace (Section III)	52M	6 prosodic features, 62.5 Hz	860 h LibriSpeech	masked prediction of prosodic tokens

the utterance’s real frames from a fixed layer), which are partitioned into $K = 30$ clusters by K-means; the cluster index $y_i \in \{1, \dots, K\}$ is the utterance’s acoustically derived prosodic style label. A text embedding model T produces sentence embeddings $T(x_i)$. H_1 is tested by (i) training classifiers to predict y_i from $T(x_i)$ on a train split and evaluating on held-out utterances, against the same classifiers trained on two reference feature sets (prediction P1), and (ii) measuring the geometry of the label partition $\{y_i\}$ directly in the text embedding space (prediction P2). Clustering is fitted once per (speaker, speech model, layer) and shared by all text models, so text-model comparisons are free of clustering noise.

B. Speech Representation Models

Table I lists the five models. The four SSL baselines span the three major self-supervised families and were the top performers of their families in the ten-model analysis of Rehman *et al.* [27]; pretrained weights are taken from the torchaudio pipelines. wav2vec2 XLSR-53 [43] additionally provides a multilingual reference, and TIMBRESpace the prosody-only control. All models are used frozen. Layers are indexed proportionally on a 1–7 scale, index i denoting the layer at $i/7$ of each model’s depth, so that a given index is the same relative depth in every model and index 7 is the final layer. The information carried by SSL layers varies strongly with depth [46], so the index is a consequential choice. All main results use the final layer. Indices 1 and 3 were also run, and Section VIII-F reports the full comparison across the three depths.

C. Text Embedding Models

Twelve text embedding models are used: similarity-trained sentence transformers (all-MiniLM-L6-v2 [39], all-mpnet-base-v2 [38], paraphrase-distilroberta-base-v2 [36], paraphrase-multilingual-MiniLM-L12-v2); general-purpose dense retrievers (nomic-embed-text-v1.5 [40], bge-large-en-v1.5, e5-large-v2); an NLI-tuned encoder (bart-large-mnli [37]); large instruction-tuned embedders (gte-Qwen2-1.5B-instruct, Qwen2-Embedding-0.6B); and two plain masked-language-model (MLM) encoders that are *not* trained for sentence similarity, roberta-base [47] and distilbert-base-uncased [48]. The last pair functions as a weak-encoder control: if the measured effect is semantic, models trained to represent sentence meaning should beat models that merely pool token embeddings.

D. Classifiers and Reference Baselines

Six scikit-learn classifiers, Logistic Regression, Random Forest, K-Nearest Neighbors, Ridge Classifier, Linear Discriminant Analysis (LDA), and Extra Trees, represent complementary geometric assumptions and guard against the known sensitivity of representation evaluation to the probing architecture [24]. Features are standardized with training statistics and the training set is subsampled to 10,000 utterances; test sets are never subsampled. All reported numbers use the *best-performing* classifier per configuration by weighted F_1 .

This selection gives each arm its most suitable probe: LDA wins 359 of 360 text-embedding cells, whereas the TF-IDF arm is won by tree ensembles in 228 cells and the length arm by Random Forest in 264. Because it is made on the held-out split, it is a maximum over six correlated estimates, and the selection gap (best minus mean over the six classifiers) of 0.030, 0.022, and 0.014 weighted F_1 respectively gives the scale of the resulting optimism in each arm’s absolute level. Every arm comparison is therefore also repeated with a single classifier held fixed across arms (Section VIII-B), which removes the selection step from the comparison.

Two reference arms run through the identical classifier suite on the same labels and the same held-out utterances:

- **Length only:** three features, the word count, character count, and heuristic syllable count of the text.
- **TF-IDF:** bag-of-words vectors, up to 2,000 features with sublinear term frequency, over scikit-learn’s default word tokens (punctuation and single-character words excluded). This arm captures “this is the cluster where the speaker says *concrete*” without any sentence-level semantics or surface form.

The length arm is the floor that any text representation must clear. The bag-of-words arm measures what word identity alone achieves; exceeding it is not required for text to be informative about delivery.

E. Confound Controls

Three properties of the pipeline could create text-predictable clusters with no prosodic style involved. Each is controlled, and the controls change the result materially.

1) *Utterance Length:* Mean-pooled speech embeddings encode input length almost perfectly: a ridge predictor recovers log utterance length with $R^2 \geq 0.97$ for every model tested. This code is structural, not prosodic. A predictor trained on natural utterances also recovers the duration of arbitrary mid-utterance crops, which contain no phrase boundary, declination arc, or final lengthening, with a predicted-versus-true slope of

1.01 to 1.12; excluding edge frames or subtracting per-position profiles does not remove it. Without intervention, K-means therefore partly returns duration bins. Two measures address this: embeddings are L2-normalized before clustering, and the length-conditional mean is removed, that is, clustering operates on $X - E[X \mid \log \ell]$, with the conditional mean fitted as a cubic spline on the training split only and applied unchanged to test data. Erasure cuts the length-only baseline’s lift by 40 to 70%. A residual association with length remains ($\eta^2 = 0.19$ to 0.52), so the length arm is retained as a diagnostic; at the final layer that residual is concentrated in the two HuBERT models (Section VIII-A).

2) *Sequence Length Beyond the Trained Range*: TIM-BRESPACE was trained on sequences of up to 622 tokens (≈ 10 s), while 43.5% of analysis sequences are longer, and whole-sequence inference on those shifted embeddings further than the distance between two unrelated utterances (median cosine 0.757 for the longest third). Long utterances are therefore encoded in overlapping 622-token windows with stride 311 and averaged per token, which leaves short sequences bit-identical and brings the model’s length sensitivity down to WavLM’s level. The waveform models were verified to be stable over the same range (cosine ≥ 0.96).

3) *Clustering Instability*: K-means at $K = 30$ is sensitive to training-row order: shuffling rows alone produced partitions with adjusted Rand index (ARI) 0.47 to 0.85 against each other. Runs therefore sort rows canonically, use float64 with ten initializations, seed all generators, and cache the fitted clustering so that all twelve text models of a (speaker, speech model) pair score the same partition. This makes every result reproducible, and every comparison between text models or arms a comparison on one fixed partition.

V. DATASETS

The analysis uses a corpus of Creative-Commons-licensed YouTube speech: approximately 1,200 hours from six speakers (three male, three female), subsampled to 200 hours per speaker to equalize speaker effects, spanning diverse topics, registers, and spontaneous conversational contexts. Per-channel attribution, hours per speaker, and the data-use statement are given in the supplementary material.

Each source video is denoised before any further processing, with a pretrained speech-enhancement model [49] applied in 300 s windows with 10 s overlap and peak-normalized, and is then transcribed with a Whisper-large-v3-turbo model using word-level timestamp alignment [50]. Denoising suppresses background noise before any embedding is computed, which narrows the recording-condition account discussed in Section IX.

Audio is segmented along transcript boundaries, merging adjacent segments separated by at most 1 s up to a 20 s maximum, and discarding segments shorter than a minimum duration, longer than the encoder’s maximum input, or carrying fewer than 5 word tokens per second, which removes near-silent and sparsely transcribed spans; the transcriber’s own silence handling means segments do not contain long pauses. This yields 22,929 to 66,619 training and 5,840 to 15,488

test segments per speaker. The text attached to a segment is its own transcript; when that is shorter than 200 characters, neighboring segments’ text is prepended and, if still short, appended, so that every text input carries at least that much context.

Train–test partitioning is at the level of source videos, drawn randomly but deterministically in an 80/20 ratio (426 to 1,270 training and 107 to 318 test videos per speaker), so no video contributes utterances to both sides and classifiers cannot exploit within-video shortcuts. Random rather than chronological selection also removes any systematic era difference between training and test data; it does not remove topic structure, since a channel’s recurring settings and subjects appear on both sides of the split.

All training and evaluation is performed on *individual speakers*: acoustic clusters are fitted independently on each speaker’s data, and text-to-cluster prediction is trained and evaluated within that same speaker’s utterances. The measured association therefore reflects how an individual varies delivery with content, not differences between speakers. The six speakers are accordingly treated as blocks, each a complete replication of the experiment, rather than as a sample of a population.

VI. EVALUATION METRICS

A. Classification Metrics and the Chance Level

Classifiers trained on text features predict acoustic cluster assignments on held-out utterances, scored by **top-3 accuracy** (the proportion of utterances whose true cluster is among the three highest-scoring predictions; the relevant metric for style selection, where proposing a small candidate set suffices) and **weighted F_1** , reported as a robustness check.

With $K = 30$ clusters, uniform chance for top-3 is $3/K = 0.10$, but K-means on speech embeddings does not produce balanced clusters (balance scores 0.10 to 0.57 here, and near zero for one degenerate encoder). The appropriate floor is therefore the **majority-top-3 rate**: the combined share of a speaker’s three largest clusters, which is the accuracy of a constant predictor that ignores the text entirely. This floor averages 0.255 across configurations and reaches 0.448 for the most unbalanced one, so an apparently large margin over uniform chance can be worth nothing. All claims below are stated against this floor.

B. Geometric Coherence Metrics

Independently of any classifier, the acoustic-cluster partition is evaluated directly in the text embedding space by the cosine **silhouette** of the acoustic clusters computed on text embeddings (positive values indicate that acoustically defined clusters are also semantically compact) and by the **separation ratio** of mean inter-cluster to mean intra-cluster distance. Raw coherence values are not comparable across text models, whose cosine similarity scales differ, so only within-model comparisons are made.

TABLE II
SPEAKER-INDEPENDENT PROBE PERFORMANCE ON IEMOCAP

Encoder	dim	Emotion		Identity	
		L1	L7	L1	L7
WavLM-Large	1024	0.570	0.586	0.984	0.949
wav2vec2-XLSR53	1024	0.580	0.578	0.986	0.966
HuBERT-XLarge	1280	0.537	0.582	0.983	0.935
HuBERT-Large	1024	0.545	0.526	0.982	0.941
TIMBRESpace	768	0.417	0.397	0.814	0.834
majority floor	—	0.197	0.197	0.692	0.692

Best weighted F_1 over the classifier set on held-out Session 5. L1 and L7 are the first and final encoder layer. The identity axis is the two unseen test speakers, who differ in gender.

C. Statistical Testing

The unit of replication is the speaker. Cells are averaged within a speaker first, so that speakers contribute equally, and one-sided Wilcoxon signed-rank tests are then applied across the six speakers; with $n = 6$ the smallest attainable one-sided p is .016 (all six speakers agreeing) and five of six gives .031. Factor effects (choice of text or speech embedding model) use Friedman tests with speakers as blocks and Kendall’s W as effect size; layer comparisons use the same cells at each depth, each scored against its own floor, with Friedman tests over depths and two-sided Wilcoxon tests between them. Per-cell binomial tests against the floor, and per-cell two-proportion tests between arms, are also reported, but they treat held-out utterances as independent Bernoulli trials; since consecutive utterances from the same video are correlated, these p -values are *nominal* and are read as descriptive indicators of effect strength. Alongside the mean over configurations, the **worst configuration per speaker** is reported: the mean describes the typical configuration, and the worst cell bounds it from below.

VII. VALIDATION ON LABELED SPEECH

The style labels this paper predicts are unsupervised by design. To verify that the embeddings behind them encode paralinguistic information that listeners label consistently, we repeat the encoding and clustering stages on IEMOCAP [51], which supplies per-utterance emotion and speaker labels, using the identical pipeline: same encoders, same pooling, same length erasure, same K-means.

We use the corpus-provided speaker-disjoint split (Sessions 1–4 for training, Session 5 for test): 3,289 training and 1,176 test segments over the four standard emotion classes, with seven distinct speakers surviving filtering in training and two entirely unseen speakers in test, so every probe below is speaker-independent. Embeddings are taken at the first and the final encoder layer. Because the test split contains one male and one female speaker, speaker identity and gender induce the same binary partition of the test set; we report it as a single *identity* axis. Supervised probes are the classifier set of Section IV-D, with the majority-class predictor as the floor.

A. Emotion Is Decodable from All Four SSL Encoders

The four SSL encoders reach 0.53 to 0.59 weighted F_1 on four-way emotion (Table II), against a majority-class floor

TABLE III
UNSUPERVISED STRUCTURE ON IEMOCAP

Encoder	Emotion $\bar{\eta}^2$			Identity $\bar{\eta}^2$		
	L1	L7	$\times$ null	L1	L7	ARI
WavLM-Large	0.056	0.055	21	0.808	0.144	0.009
wav2vec2-XLSR53	0.056	0.035	14	0.470	0.019	0.007
HuBERT-Large	0.058	0.036	14	0.550	0.039	0.013
HuBERT-XLarge	0.049	0.041	16	0.367	0.059	0.014
TIMBRESpace	0.086	0.166	69	0.169	0.161	0.024

Mean $\bar{\eta}^2$ over ten UMAP components with no labels fitted; “ $\times$ null” is the L7 emotion $\bar{\eta}^2$ as a multiple of its expected value under H_0 . ARI: adjusted Rand index between the K-means partition ($K = 30$) and the emotion label at the final layer.

of 0.197 (0.367 accuracy; class counts 432/323/260/161), clearing that floor by 16 to 22 accuracy points in every cell ($p \leq 1.4 \times 10^{-27}$, binomial). The representations the pipeline clusters are not clustering on nothing.

These values sit below the accuracies above 70% commonly reported for IEMOCAP [31], as expected: published numbers are typically obtained by fine-tuning the encoder, pooling across layers, and retaining utterance duration, whereas ours come from a single frozen layer with the *length confound erased* (Section IV-E), and duration is a genuine paralinguistic cue in acted corpora, not only a confound. At the first layer, where the text baselines were logged on this corpus, a length-only baseline reaches 0.111 to 0.195 top-3 against a majority-class floor of 0.216 to 0.283, that is, *below* a constant predictor for all five encoders, as on the main corpus at the same depth. Table II is therefore a conservative lower bound on the style information available to the pipeline, not a speech-emotion-recognition benchmark.

B. Identity Dominates the Embedding Geometry

At the first layer the identity axis is an order of magnitude stronger than the style axis in every SSL space (Table III): a *single* unsupervised component separates the two held-out speakers at $\bar{\eta}^2 = 0.71$ to 0.95, while no component exceeds $\bar{\eta}^2 = 0.12$ for emotion, and K-means purity against identity reaches 0.81 to 0.91 with no supervision at all. This is the empirical basis for the strict within-speaker protocol: pooling speakers before clustering yields identity partitions, not style partitions. The clustering in Table III is deliberately cross-speaker for that reason; it is the contrast case, not the method.

C. The Non-Lexical Control Behaves as Required

TIMBRESpace is the weakest encoder under supervised probing, which is expected: 768 dimensions derived from six acoustic streams with no phonetic channel cannot compete with 1024- to 1280-dimensional waveform encoders at recovering categorical labels. What matters for its role is whether its space is *organized* along paralinguistic axes, and under unsupervised measurement it is the most strongly organized of the five. At the final layer it leads every label-free metric on both axes: mean $\bar{\eta}^2$ at 69 times the null, three times the strongest SSL encoder and 4.7 times the weakest, maximum $\bar{\eta}^2$ of 0.248, cluster-label ARI and adjusted mutual information,

and cluster–emotion $\chi^2 = 506$ ($p < 10^{-40}$), while separating the two unseen speakers at $F_1 = 0.83$ from acoustic features alone. Because the pipeline consumes these embeddings without supervision, this is the property it actually depends on.

D. Layer Choice Changes Geometry More Than Decodability

Between the first and final layer, SSL probe performance is essentially flat (within $0.05 F_1$), but the underlying geometry is not: identity $\bar{\eta}^2$ falls by 6 to 25 times and the cluster–identity ARI by 4 to 20 times, as the top layer specializes for the masked-prediction objective. TIMBRESpace is stable or improving across layers, and its emotion $\bar{\eta}^2$ doubles at the final layer. A supervised probe is therefore a poor instrument for selecting a layer for a clustering pipeline, and the unsupervised side does not settle the choice either: emotion $\bar{\eta}^2$ is equal or slightly higher at the first layer for the four SSL encoders, whereas at the final layer their geometry is far less dominated by voice identity and the prosody-only control is at its most emotion-aligned. One encoder is degenerate at the final layer: ten principal components carry 91.8% of XLSR-53’s variance, against 33 to 54% for the others, so its tight clusters (silhouette +0.18) sit in a nearly low-rank space. The same collapse recurs on the main corpus. Because labeled speech does not decide between depths, the choice for the main experiments is made on the main corpus itself (Section VIII-F).

VIII. RESULTS

The experiment comprises 360 configurations, 6 speakers $\times$ 5 speech models $\times$ 12 text models, all complete. Every configuration uses $K = 30$, the final encoder layer, and length-erased clustering. At the final layer one speech model, XLSR-53, yields a degenerate partition (Section VIII-D). It is kept in every mean, since the majority floor scores its imbalance correctly, and results without it are given wherever they change a conclusion.

A. Does Text Predict the Acoustic Cluster at All?

All three arms exceed uniform chance in 360 of 360 cells, which illustrates why that reference is uninformative here: so does a constant predictor. Against the majority-top-3 floor (Fig. 2, Table IV), the text embedding arm averages 0.366 top-3 against a floor of 0.255, a lift of +0.111, and exceeds the floor in 293 of 360 cells (276 significantly after Holm correction). At the speaker level it clears the floor for all six speakers (Wilcoxon $p = .016$). The margin is smallest for M1 (+0.043), whose clusters are the most unbalanced (floor 0.309). The cells that fail are concentrated in one encoder: XLSR-53 clears its floor significantly in 10 of 72 cells, the other four speech models in 50 to 72 of 72.

The length-only arm averages 0.254, which is *at* the floor (mean lift -0.001), exceeds it for three of six speakers, and is significantly above it in 120 of 360 cells. Those cells are concentrated in the two HuBERT encoders (60 and 48 of 72; mean lift +0.035 and +0.031) and WavLM (12 of 72); for XLSR-53 and TIMBRESpace the length arm never significantly exceeds the floor. At the first layer the same

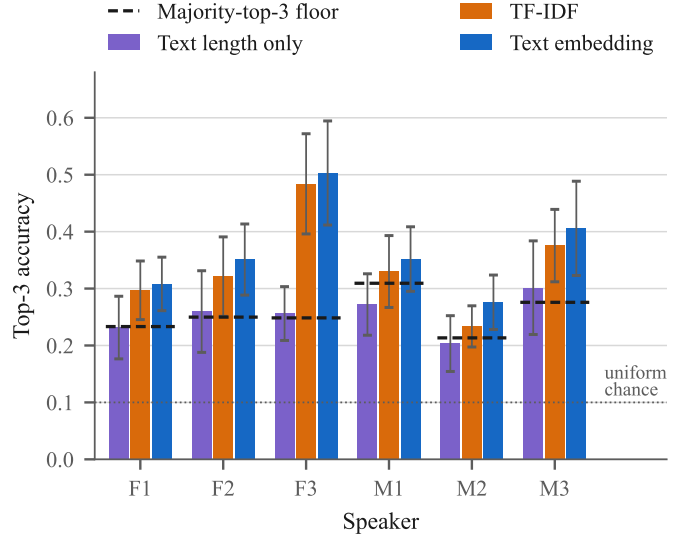


Fig. 2. Top-3 accuracy of the three text arms per speaker, averaged over the twelve text models and five speech models (error bars: standard deviation over those cells). The dashed rule is each speaker’s majority-top-3 floor; uniform chance (0.10) is shown for comparison only.

TABLE IV
PER-SPEAKER TOP-3 ACCURACY OF THE THREE TEXT ARMS

Spk	Floor	Len	TF-IDF	Sem	Sem-TFIDF	worst
F1	0.233	0.231	0.297	0.308	+0.011	−0.058
F2	0.250	0.260	0.321	0.351	+0.030	−0.037
F3	0.248	0.256	0.484	0.503	+0.019	−0.038
M1	0.309	0.272	0.330	0.352	+0.022	−0.040
M2	0.213	0.203	0.234	0.276	+0.042	+0.004
M3	0.276	0.302	0.375	0.406	+0.030	−0.020
mean	0.255	0.254	0.340	0.366	+0.026	

Averaged over the twelve text models and five speech models. Floor: majority-top-3 rate. Sem: text embedding arm. worst: minimum of Sem-TFIDF over that speaker’s configurations. Wilcoxon across the six speakers: Sem > floor, $p = .016$; Sem > TF-IDF, $p = .016$; worst-case Sem > TF-IDF, $p = .98$.

arm sat *below* the floor for all six speakers (Section VIII-F), so length-conditional erasure is less complete at the final layer of the HuBERT models, where the residual amounts to about a fifth of the embedding arm’s lift. The embedding arm nonetheless exceeds the length arm in 359 of 360 cells, and for every speaker under each of the three linear probes.

B. What Carries the Association?

The bag-of-words arm answers this directly. TF-IDF averages 0.340 top-3, a lift of +0.085 over the floor, which is 77% of the text embedding arm’s lift measured as a ratio of per-speaker means (median per-cell share 86%; 67% excluding speaker F3). Sentence embeddings add +0.026 top-3 over TF-IDF on average (+0.025 weighted F_1), positive in 76% of cells (97% on weighted F_1) and for all six speakers ($p = .016$ on both metrics). Per cell, the embedding arm beats TF-IDF significantly in 187 of 360 configurations and loses significantly in 18.

Two observations show how thin this increment is. First, it is not uniform across configurations: in each speaker’s *worst*

TABLE V
INCREMENT OF TEXT EMBEDDINGS OVER BAG-OF-WORDS WITH THE CLASSIFIER HELD FIXED

Fixed classifier	All 12			MLM	Sim.
	Sem-TFIDF	Spk > 0	p		
K-Nearest Neighbors	+0.066	6/6	.016	+0.080	+0.063
Linear Discriminant	+0.032	6/6	.016	+0.051	+0.028
Logistic Regression	+0.020	6/6	.016	+0.037	+0.017
Ridge Classifier	+0.019	6/6	.016	+0.036	+0.015
Random Forest	−0.013	0/6	1.00	+0.014	−0.020
Extra Trees	−0.017	0/6	1.00	+0.010	−0.024
best per configuration	+0.025	6/6	.016		

Each row holds one classifier fixed across the embedding and TF-IDF arms (weighted F_1 , averaged within speaker; one-sided Wilcoxon across the six speakers). MLM: roberta-base and distilbert-base-uncased. Sim.: the seven similarity-trained sentence encoders and the two instruction-tuned embedders. Last row: the best-on-test selection used elsewhere in this paper.

configuration, the embedding arm beats TF-IDF for only one of six speakers (M2, +0.004; Wilcoxon $p = .98$), and on weighted F_1 for four of six. Second, its sign depends on the probe and on the encoder (Table V). Under each linear probe the embedding arm beats TF-IDF for every speaker. Under both tree ensembles, however, the nine similarity-trained and instruction-tuned encoders lose to TF-IDF for every speaker, because a tree can split on an individual informative word, which is exactly the mechanism the bag-of-words arm was built to capture. The nearest-neighbor row is not evidence either way, since it is a poor probe for sparse high-dimensional TF-IDF. For those nine encoders the defensible statement is therefore that their embeddings are more *linearly* separable with respect to the cluster labels than bag-of-words vectors are, not that they carry more information about them. The exception is the two MLM encoders, whose advantage survives both tree ensembles (+0.014 and +0.010, for six and five of six speakers): the only increment over bag-of-words that is robust to the choice of probe comes from encoders that retain token identity and surface form.

Across cells, embedding accuracy tracks bag-of-words accuracy almost exactly (Spearman $\rho = 0.92$ over cells, 0.95 over the 30 speaker–speech pairs): wherever text predicts the clusters, it does so mostly through which words were used. Speaker F3 is the clearest case. It has by far the highest predictability of any speaker (0.503 top-3), but TF-IDF alone reaches 0.484, and on weighted F_1 the embedding arm is only marginally above it (0.260 against 0.245). Its clusters track vocabulary, an observation taken up in Section IX.

C. Does Better Semantics Help?

If the residual beyond word identity were semantic, encoders trained to represent sentence meaning should extract more of it than encoders that merely pool token embeddings. The opposite holds (Fig. 3). roberta-base (+0.063) and distilbert-base-uncased (+0.051) lead every other encoder and rank first and second for each of the six speakers individually; roberta-base beats every non-MLM encoder in at least 29 of the 30 speaker–speech blocks. The seven

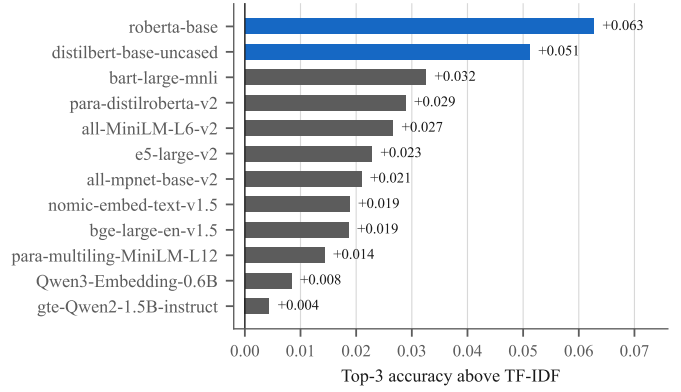


Fig. 3. Top-3 accuracy above the TF-IDF baseline for each text embedding model, averaged over speakers and speech models. Highlighted: the two MLM encoders, included as weak-encoder controls. para-multiling-MiniLM-L12 and para-distilroberta-v2 abbreviate paraphrase-multilingual-MiniLM-L12-v2 and paraphrase-distilroberta-base-v2.

similarity-trained sentence encoders fall between +0.014 and +0.029 and the NLI-tuned bart-large-mnli at +0.032, while the 0.6B and 1.5B instruction-tuned embedders rank last (+0.008 and +0.004). The text-model effect is itself significant (Friedman over the twelve encoders, $\chi^2(11) = 48.5$, $p = 1.2 \times 10^{-6}$, $W = 0.74$ for top-3; $\chi^2(11) = 48.9$, $p = 1.0 \times 10^{-6}$, $W = 0.74$ for weighted F_1), so this ordering is not noise.

The natural reading is that the residual lives in *surface form*, such as word order, function words, punctuation, disfluency, and sub-word composition, which raw MLM token embeddings retain and similarity-trained encoders deliberately normalize away; it shrinks rather than grows with model scale, and Table V shows that only the MLM encoders keep their advantage under probes that can split on single words. Part of that surface form may be prosodically meaningful in its own right. The bag-of-words arm discards punctuation (Section IV-D), so a question mark that the MLM encoders see is invisible to it; their advantage therefore argues against a *semantic* reading of the increment, not against a prosodic one.

D. Speech Models and the Prosody-Only Control

Prediction P3 states that the association should persist for a speech representation with no lexical channel. It persists, but weakened (Figs. 4 and 5). The three working SSL waveform encoders reach 0.365 to 0.419 top-3, a lift of +0.157 to +0.174 over their floors, while TIMBRESpace reaches 0.274, a lift of +0.070 that is significant in 50 of its 72 configurations but below each of the three waveform encoders for every one of the six speakers (Friedman over the four encoders, $\chi^2(3) = 11.6$, $p = .009$). The increment over TF-IDF is also smallest for TIMBRESpace (+0.010, against +0.026 to +0.035 for the waveform encoders) and negative for two speakers, although at this depth the increments do not differ significantly across speech models (Friedman $p = .50$).

XLSR-53 is a separate case. Its final layer is nearly low-rank: ten principal components carry 92% of its variance, against 43 to 57% for the other encoders, and K-means returns

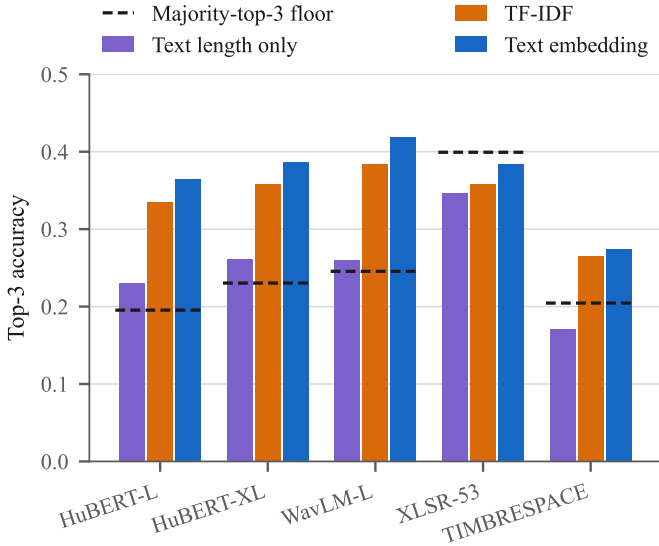


Fig. 4. Top-3 accuracy of the three text arms per speech representation model, averaged over speakers and the twelve text models, against the corresponding majority-top-3 floor (dashed).

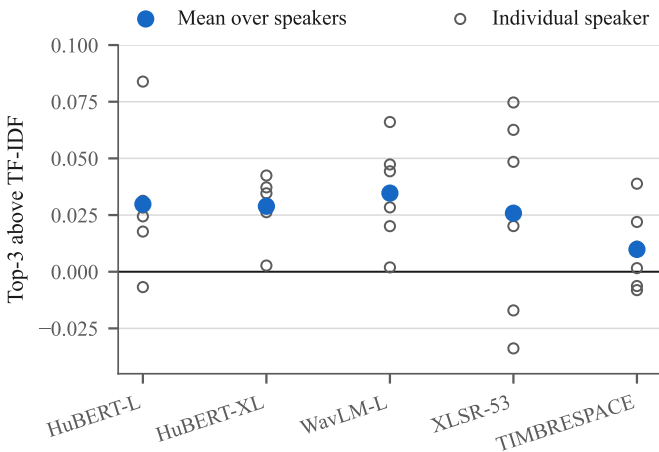


Fig. 5. Top-3 accuracy above TF-IDF per speech representation model. Open circles are individual speakers (mean over the twelve text models); filled circles are the mean over the six speakers.

a few large, tightly separated clusters (acoustic silhouette $+0.29$, against -0.08 to -0.05 elsewhere; balance 0.01), the three largest holding 40% of held-out utterances. No text arm beats that floor on average (embedding-arm lift -0.015), although the embedding arm still exceeds TF-IDF ($+0.026$). The same collapse appears at the final layer on IEMOCAP (Section VII-D) and is absent at the two earlier depths (Section VIII-F), so it is a property of this checkpoint’s final layer rather than of the corpus or the pipeline. Its well-separated but text-unpredictable clusters also show that cluster compactness and text predictability are distinct properties.

The weakening is not uniform across speakers. For five of the six, TIMBRESpace retains between none and about a third of the three waveform encoders’ lift, the pattern expected if lexical content leaks from the speech embeddings into the cluster labels. The sixth, F3, whose association is the largest in the corpus, does not fit this account. The prosody-only

representation retains 81% of the waveform encoders’ lift for F3 ($+0.273$ against $+0.338$), bag-of-words alone predicts its clusters at $+0.262$ above the floor, and the ratio is 0.69 to 0.78 at every depth of Section VIII-F. Structure that text predicts in a representation without a phonetic channel cannot be lexical leakage through the speech encoder. It must be carried by pitch, formant, energy, or SNR differences: either by delivery, or by recording conditions, which the two channel-sensitive features, power and SNR, would register (Section IX). As a lexical-leakage control, TIMBRESpace therefore gives a speaker-dependent answer: leakage can account for most of the association in five speakers, but not in the speaker who contributes most of it.

The gap between TIMBRESpace and the waveform encoders also co-varies with model capacity. Across the 30 speaker–speech pairs, the lift over the floor correlates with the width of the speech embedding ($\rho = 0.38$ speaker-centered, 0.45 without XLSR-53, against 0.77 at the first layer), and TIMBRESpace, the narrowest encoder (768 dimensions against 1024 and 1280), also produces the least compact clusters, with roughly half the Calinski–Harabasz score of the working SSL models. Wider encoders yield more separable clusters, which are easier to predict from any input, so the weaker average association of the prosody-only representation may in part reflect that narrower, less separable clusters are harder to predict. Pretraining domain does not explain the gap: the two HuBERT models, which show among the strongest associations, were pretrained on read audiobook speech, as TIMBRESpace was. The compactness account is not the whole story either: on labeled speech the same encoder produces the least compact clusters and yet the *best* alignment with emotion and speaker (Section VII), so low cluster quality does not by itself imply a representation with little recoverable structure.

E. Cluster Geometry

Prediction P2 is not supported. The cosine silhouette of the acoustic clusters measured in text embedding space is negative in *all* 360 configurations (range -0.230 to -0.025), with a mean separation ratio of 0.253, above one in only 6 of 360: utterances sharing an acoustic cluster are no closer in text space than utterances from different clusters. This holds even though the coherence metrics are computed on the most favorable subsample available, the at most 100 utterances nearest each cluster centroid. A classifier can still extract a weak signal from systematic differences in mean location, but the clusters are not “about” distinct content or registers in any geometric sense.

The geometry is not uninformative, though. While the silhouette is, if anything, inversely related to predictability ($\rho = -0.34$ across the 30 speaker–speech pairs), the *separation ratio* tracks it ($\rho = 0.61$, and 0.58 after centering within speakers). As a reference level, the clusters are only weakly formed in the acoustic space that defines them (mean silhouette -0.08 to -0.05 for the four non-degenerate encoders), so compactness is a demanding criterion for a 30-way partition in either modality.

TABLE VI
SENSITIVITY TO ENCODER DEPTH

Layer index	1	3	7	p
Majority-top-3 floor	0.219	0.220	0.255	.011
Embedding-arm top-3	0.343	0.361	0.389	.030
Lift over floor	0.124	0.141	0.134	.042
<i>excluding XLSR-53</i>				
Floor	0.217	0.221	0.219	.61
Lift over floor	0.125	0.143	0.169	.030
Weighted F_1	0.153	0.164	0.178	.002
Speakers with lift > 0	6/6	6/6	6/6	
Speaker share of variance	0.87	0.75	0.52	
Text-space separation ratio	0.379	0.417	0.472	.006
TF-IDF share of lift	0.71	–	0.64	
Length-arm lift	−0.022	–	−0.001	
Sem-TFIDF, weighted F_1	+0.024	–	+0.039	.016

The three text models run at every depth (roberta-base, distilbert-base-uncased, bart-large-mnli): 90 matched configurations per depth, each scored against the majority-top-3 floor of its own clustering. p : Friedman test over the three depths with speakers as blocks; last row, one-sided Wilcoxon of index 7 against index 1. The baselines were not run at index 3.

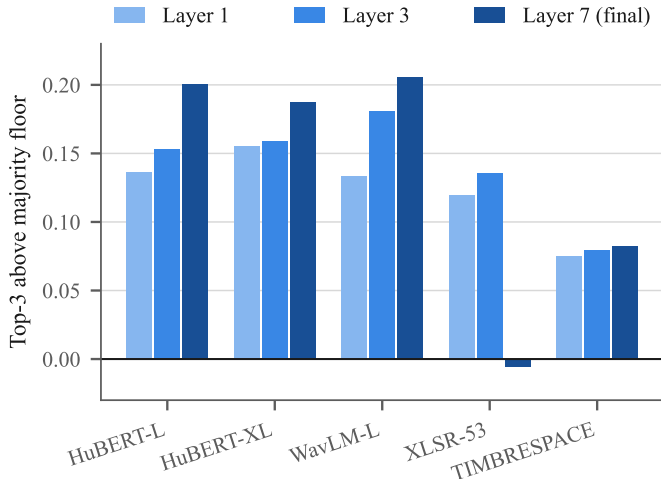


Fig. 6. Top-3 accuracy above each configuration’s majority floor, per speech representation model and encoder depth (index 1 is the first seventh of the model, 7 the final layer), averaged over the six speakers and the three text models of Table VI.

F. Sensitivity to Encoder Depth

The final layer was fixed after indices 1 and 3 had also been run, so all three are reported. Table VI compares them over matched configurations for the three text models run at every depth, which are also the three strongest at the final layer, and Fig. 6 breaks the lift down by speech model.

Excluding the collapsed XLSR-53, the final layer gives the largest lift over the floor (Friedman $p = .030$) and the largest weighted F_1 for all six speakers ($p = .031$ against either earlier depth), while the floor itself does not move, so the gain is not an imbalance artifact. Per encoder, the final layer gives the largest lift for five of six speakers for each of HuBERT-L, HuBERT-XL, WavLM-L and TIMBRESpace; XLSR-53 peaks at index 3 and collapses at the final layer. The association is also less concentrated in one speaker at depth: the speaker share of variance in top-3 accuracy falls from 0.87

to 0.52, and the weakest speaker’s lift rises from +0.016 to +0.074.

The qualitative conclusions do not depend on depth. At every depth the embedding arm clears the floor for all six speakers, and the acoustic clusters have a negative silhouette in text space in every configuration. Where the baselines were run, the embedding arm exceeds TF-IDF for all six speakers at both depths, and roberta-base and distilbert-base-uncased head the text-model ranking at both. Two properties do change with depth: length erasure is more complete at the first layer, and XLSR-53 is usable only below its final layer. The final layer is used because it is where the association that H_1 predicts is strongest for four of the five encoders, so negative findings at that depth cannot be attributed to an unfavorable choice of layer. P1, the positive result, is thereby reported at its most favorable depth, and it holds at every depth examined.

G. Verdict on the Null

H_0 is rejected, and what replaces it is a small, lexically carried association. P1 is satisfied in a weak form: all six speakers clear the majority floor ($p = .016$), three quarters of that association is available from word identity alone, and the embedding arm’s advantage over bag-of-words, although positive for all six speakers, holds for similarity-trained encoders only under linear probes. The association varies more across speakers than across models. Speaker identity accounts for 55% of the variance in top-3 accuracy across the grid, against 25% for the speech model and 3% for the text model; one speaker (F3) contributes a lift of +0.255 against +0.082 averaged over the other five; and in each speaker’s worst configuration the lift over the floor is negative for all six, in every case an XLSR-53 configuration, and for two of six when that encoder is set aside. P2 fails in every configuration at every depth, and P3 holds only in part (Section VIII-D). The hypothesis that text alone identifies prosodic style is therefore supported in a bounded form: text is informative about these utterance-level delivery clusters, mostly through which words are used, and sentence-level meaning adds little.

IX. DISCUSSION

A. A Lexically Mediated Association

The association that survives all three controls is real but narrow, and its character is not what the TTS literature assumes: it is carried by vocabulary. Two accounts of that vocabulary effect are consistent with the results of Section VIII, and they differ in what a TTS system could do with it. In the first, a speaker’s vocabulary co-varies with the situation in which they are recorded: what someone talks about changes with where, when, and how they are recorded, and the acoustic embedding registers the accompanying differences in vocal effort, distance, room, and microphone. The text side then predicts a cluster without any prosodic mediation. Denoising eliminates the crudest version of this channel story; microphone timbre, reverberation, distance, and vocal effort survive it, and RMS power and SNR are two of the prosody-only model’s six features. The dominant speaker fits this account:

F3’s clusters are almost as text-predictable in a space of pitch, formant, energy, and SNR features as in the waveform encoders.

In the second account, words are genuine cues to delivery. Interrogatives, intensifiers, discourse markers, and punctuation signal the register in which a stretch of speech is produced, so an association carried by word identity would be prosodic coupling through the lexicon rather than a confound, and would be usable by a TTS system as it stands. The accounts make different predictions about which words carry the association, content words, which track topic, or function words, discourse markers, and punctuation, which cue delivery, and about how strongly each speaker’s clusters align with source video; both are the natural next tests within this protocol. The results do not establish that prosody is unpredictable from text. They establish that at the granularity of this design, unsupervised 30-cluster partitions of mean-pooled within-speaker utterance embeddings, and with trivial explanations controlled, text informs delivery mostly through which words are used.

B. The Controls Are the Finding

Without the controls of Section IV-E, the same data would have supported a strong positive result, and each control removes a specific share of it. Pooled speech embeddings encode input length structurally, so clustering them partly returns duration bins that text length predicts trivially; after erasure the length arm falls to the level of a constant predictor. The control is conservative: duration is a genuine paralinguistic cue as well as a confound, so erasing it biases every arm towards the null. The majority floor absorbs more than half of the apparent margin over uniform chance: the collapsed final layer of XLSR-53 reaches 0.384 top-3, nearly four times uniform chance, without beating a constant predictor. The bag-of-words arm separates “text is associated with delivery” from “sentence meaning is associated with delivery”: without it, the 0.111 lift over the floor would be attributed to semantics, when three quarters of it is word identity. We recommend all three as minimum practice for claims of this type, whatever their sign.

C. Implications for Reference-Free TTS

A retrieval mechanism that selects prosodic parameters by matching input text against the transcriptions of each style cluster would, in this corpus, operate at roughly 0.37 top-3 against a 0.26 constant-predictor floor, with most of that margin carried by vocabulary. To the extent that vocabulary reflects topic rather than delivery, such a mechanism would retrieve the styles a speaker uses when discussing similar subjects, which is not the same as the style the sentence calls for. This bounds, rather than contradicts, the systems that exploit the text-to-style link inside a synthesis pipeline [6], [8], [10]: they co-train the style representation with the objective, which may recover structure that unsupervised acoustic clustering does not expose, and their gains are measured by listener preference, which a weak but non-zero signal can still improve. What the result does undercut is the expectation that scale and data alone will deliver style selection from text [14], [15]: the largest text encoders extracted the least beyond bag-of-words.

D. Limitations

The claims above are bounded by the design in the following ways. *Labels*. The style labels are unsupervised acoustic partitions, not perceptually validated speaking styles. Section VII establishes that the embeddings being clustered carry decodable paralinguistic information and that the prosody-only space is organized along it, while categorical emotion accounts for little of a 30-way partition (cluster–emotion ARI at most 0.024). What the clusters index beyond that, delivery or recording context, is the question of Section IX; per-cluster acoustic profiling and listening tests would address it. *Clustering*. Each (speaker, speech model) pair uses one clustering at $K = 30$, shared by all text models and arms, so comparisons among them are made on identical labels, whereas absolute levels would vary across seeds (ARI 0.47 to 0.85 between reorderings). *Probes*. The classifier is selected on the held-out split, so absolute levels are optimistic on the scale given in Section IV-D, and the central comparison is reported under fixed probes for that reason. *Speech models*. The speech-model comparison is not matched for embedding width or cluster compactness (Section VIII-D). *Corpus*. Each speaker is one YouTube channel, which the within-speaker design holds fixed, but recording situation varies within a channel and recurring topics and settings appear on both sides of the split. Transcriptions are ASR-generated, so every text arm follows the transcriber’s punctuation and disfluency conventions, ASR error patterns co-vary with acoustic conditions, and the bag-of-words arm differs from the embedding arms in access to punctuation as well as to sentence-level representation. All data are English conversational YouTube speech from six speakers, analyzed as blocks; read speech, other languages, and population-level inference are outside the scope of this study.

These bounds notwithstanding, the test is made under the conditions of practical use. The corpus is spontaneous, in-the-wild speech spanning many topics, settings, and recording situations, which is the material a deployed reference-free system has to handle, and a text-to-style link robust enough to be relied upon should be measurable under such conditions and not only in controlled read speech.

X. CONCLUSION

This paper turned a foundational assumption of expressive TTS, that an appropriate prosodic style can be inferred from text alone, into a falsifiable hypothesis with an explicit null, and tested it against independently derived, acoustics-only style labels under three controls: length erasure before clustering, a majority-class floor in place of uniform chance, and a bag-of-words reference arm. Across 360 configurations spanning six speakers, five speech representation models including a prosody-only control, twelve text embedding models, and six classifier families, with three encoder depths compared, the result is a weak, lexically mediated association. Text predicts its own utterance’s acoustic cluster above the majority floor for all six speakers (+0.111 top-3), but bag-of-words alone delivers about three quarters of that lift. Sentence embeddings add +0.026, an increment that changes sign with the probe for

every similarity-trained encoder and is largest for encoders not trained for sentence semantics. The association is weaker in the prosody-only representation for five of six speakers, but not for the speaker in whom it is strongest. Acoustic clusters are not compact in text embedding space in any configuration, speaker identity accounts for over half of the variance in the headline metric, and these conclusions hold at each encoder depth at which they could be tested.

Within this design, then, text informs utterance-level delivery mainly through which words are used. Whether those words act as cues to delivery or as markers of topic and recording situation is the question this result opens, and reference-free style selection cannot assume more than this. The protocol extends directly to the analyses that would answer it: separating content words from delivery cues in the bag-of-words vocabulary, relating clusters to source video, and perceptual validation of the clusters. Further extensions are listed in the supplementary material.

REFERENCES

- [1] Y. Wang *et al.*, “Tacotron: Towards end-to-end speech synthesis,” in *Proceedings of Interspeech*, 2017, pp. 4006–4010.
- [2] J. Shen *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4779–4783.
- [3] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.
- [4] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 19 594–19 621, 2023.
- [5] M. Łajszczak, G. Cámara, Y. Li, F. Beyhan, A. Van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski *et al.*, “Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data,” *arXiv preprint arXiv:2402.08093*, 2024.
- [6] D. Stanton, Y. Wang, and R. Skerry-Ryan, “Predicting expressive speaking style from text in end-to-end speech synthesis,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.
- [7] S. Lei, Y. Zhou, L. Chen, Z. Wu, S. Kang, and H. Meng, “Context-aware coherent speaking style prediction with hierarchical transformers for audiobook speech synthesis,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [8] D. Guo, X. Zhu, L. Xue, Y. Zhang, W. Tian, and L. Xie, “Text-aware and context-aware expressive audiobook speech synthesis,” *Interspeech* 2024, 2024.
- [9] S. Inoue, K. Zhou, S. Wang, and H. Li, “Hierarchical emotion prediction and control in text-to-speech synthesis,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 601–10 605.
- [10] Z. Ye, R. Huang, Y. Ren, Z. Jiang, J. Liu, J. He, X. Yin, and Z. Zhao, “Clapspeech: Learning prosody from text context with contrastive language-audio pre-training,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 9317–9331.
- [11] X. Li, S. Liu, M. W. Y. Lam, Z. Wu, C. Weng, and H. M. Meng, “Diverse and expressive speech prosody prediction with denoising diffusion probabilistic model,” in *Interspeech*, 2023.
- [12] H.-S. Oh, S.-H. Lee, and S.-W. Lee, “Diffprosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2654–2666, 2024.
- [13] T. Wang, Z. Ma, Y. Peng, H. Wang, Z. Niu, Z. Huang, Y. Wu, Y.-W. Chao, Y. Jiang, Y. Lu *et al.*, “Evaluating the expressive appropriateness of speech in rich contexts,” in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026, pp. 9088–9106.
- [14] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [15] R. S. Sutton, “The bitter lesson,” <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>, mar 2019, blog post, accessed 2026-09-20.
- [16] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.
- [17] J. J. Gumperz, “Contextualization and understanding,” in *Rethinking context: Language as an interactive phenomenon*. Great Britain: Cambridge University Press, 1992, pp. 229–252.
- [18] J. Cole, “Prosody in context: a review,” *Language, Cognition and Neuroscience*, vol. 30, no. 1-2, pp. 1–31, 2015.
- [19] S. Yang *et al.*, “Superb: Speech processing universal performance benchmark,” in *Proceedings of Interspeech*, 2021, pp. 1194–1198.
- [20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [21] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [22] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [23] M. de Seyssel, M. Lavechin, H. Titeux, A. Thomas, G. Virlet, A. S. Revilla, G. Wisniewski, B. Ludusan, and E. Dupoux, “Prosaudit, a prosodic benchmark for self-supervised speech models,” in *INTERSPEECH 2023*. ISCA, 2023, pp. 2963–2967.
- [24] S. Zaïem, Y. Kemiche, T. Parcollet, S. Essid, and M. Morchid, “Speech self-supervised representations benchmarking: A case for larger probing heads,” *Computer Speech & Language*, vol. 84, p. 101563, 2024.
- [25] T. Ashihara, M. Delcroix, T. Moriya, K. Matsuura, T. Asami, and Y. Ijima, “What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 10 166–10 170.
- [26] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, “Contentvec: An improved self-supervised speech representation by disentangling speakers,” in *Proceedings of the 39th International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 18 003–18 017.
- [27] A. Rehman, K. Jayathunge, J.-J. Zhang, and X. Yang, “Speech AI models learn boundaries, not patterns: An analysis of supervised vs. unsupervised capabilities,” in *Conference on Speech Technology and Human-Computer Dialogue (SpED)*. IEEE, 2025.
- [28] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.
- [29] A. Łańcucki, “Fastpitch: Parallel text-to-speech with pitch prediction,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.
- [30] Y. Hu, C. Zhang, J. Shi, J. Lian, M. Ostendorf, and D. Yu, “Prosodybert: Self-supervised prosody representation for style-controllable tts,” 2022.
- [31] Z.-T. Liu, M.-T. Han, B.-H. Wu, and A. Rehman, “Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning,” *Applied Acoustics*, vol. 202, p. 109178, 2023.
- [32] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, “A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical parametric speech synthesis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 157–170, 2019.
- [33] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.
- [34] Z. Zhang, H. Li, J. Hu, W. Li, and L. Xie, “Histyle: Hierarchical style embedding predictor for text-prompt-guided controllable speech synthesis,” in *National Conference on Man-Machine Speech Communication*. Springer, 2025, pp. 522–535.

- [35] E. Eren, Q. Liu, H. Kim, P. Garrido, and A. Alwan, "Promode: A speech prosody model conditioned on acoustic and textual inputs," *Interspeech 2025*, 2025.
- [36] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: <http://arxiv.org/abs/1908.10084>
- [37] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.
- [38] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in neural information processing systems*, vol. 33, pp. 16 857–16 867, 2020.
- [39] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in neural information processing systems*, vol. 33, pp. 5776–5788, 2020.
- [40] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar, "Nomic embed: Training a reproducible long context text embedder," *arXiv preprint arXiv:2402.01613*, 2024.
- [41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
- [42] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [43] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Un-supervised cross-lingual representation learning for speech recognition," *Interspeech 2021*, 2021.
- [44] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 2009, pp. 41–48.
- [45] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [46] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What do self-supervised speech models know about words?" *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 372–391, 2024.
- [47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [48] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [49] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.
- [50] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [51] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.