

Echoes in Plato’s Cave: Measuring Global and Local Alignment Between Speech and Language Representations

Alef Iury Siqueira Ferreira¹, Lucas Rafael Stefanel Gris¹, Frederico Santos de Oliveira²,
Arlindo Rodrigues Galvão Filho¹, Anderson da Silva Soares¹

¹Federal University of Goiás, Goiânia, Brazil

²Federal University of Mato Grosso, Cuiabá, Brazil

Correspondence: alef_iury_c.c@discente.ufg.br

Abstract

We investigate representational alignment between nine independently trained speech encoders and ten text encoders. We build a controlled corpus in which the same sentences are rendered by different speakers and under different emotional conditions, so that prosody varies while linguistic content is held fixed. We compare every speech-text layer pair and calibrate the strongest match with a permutation procedure that prevents the search over pairs from inflating alignment scores. Local neighborhood agreement between modalities survives calibration throughout and is strongest between late speech layers and early text layers, providing strong evidence of genuine local alignment. Global similarity measures, however, do not provide reliable evidence that the two modalities share a common geometric space as a whole. The magnitude of the alignment is unaffected by speaker identity and prosodic variation, so the shared structure is tied to the content of the sentence itself rather than to how it is spoken. It also does not increase with model capability or with how well a text model reproduces human judgments of meaning, suggesting that this shared structure corresponds more closely to lexical and surface-level representations than to fine-grained sentence meaning. These findings indicate where speech and text representations can be connected, suggest that cross-modal transfer may benefit from exploiting neighborhood structure rather than assuming a shared global geometry, and provide new alignment methods with an existing lexical correspondence to build on, while pointing to sentence-level meaning as the structure that future speech representations could capture so that alignment extends beyond the lexical level.

1 Introduction

The development of foundation models across language, vision, and audio has produced systems capable of generalizing across a wide range of tasks

and domains (Bommasani et al., 2021; Radford et al., 2021; Girdhar et al., 2023; Baevski et al., 2020; Chen et al., 2022b). Despite differences in their architectures, training objectives, and input modalities, their internal representations can still share similar structures (Huh et al., 2024). The Platonic Representation Hypothesis (PRH) interprets this similarity as evidence that increasingly capable models converge toward a shared statistical representation of reality (Huh et al., 2024). Recent work, however, shows that this conclusion depends strongly on how similarity is measured. After accounting for the effects of representation width and the search across multiple layer pairs, evidence for shared global geometry weakens, while agreement between local neighborhoods remains significant, motivating the Aristotelian Representation Hypothesis (ARH) (Gröger et al., 2026; Koepke et al., 2026).

Whether this distinction also holds between spoken and written language remains unclear. Previous work on cross-modal convergence has focused primarily on vision-text representations, while studies that include audio typically rely on general audio-caption pairs processed by a single general-purpose encoder (Huh et al., 2024; Gröger et al., 2026; Koepke et al., 2026). Speech and text have been compared directly, but inside jointly trained multi-modal models and end-to-end Speech Large Language Models (Lee et al., 2024; Hsu et al., 2026), where alignment is imposed explicitly by training rather than emergent. Because speech encoders distribute linguistic, speaker, and prosodic information differently across depth (Pasad et al., 2021, 2023; Chiu et al., 2025; Mohamed et al., 2024; de la Fuente and Jurafsky, 2024), alignment between independently trained speech and text encoders may depend on both the representations compared and the acoustic characteristics of the utterance.

In this work, we investigate whether speech and language models share a common global represen-

tational geometry or align only at the level of local neighborhoods. Because PRH makes predictions about how alignment should change with model capability and how it should behave under nuisance variation, testing it requires many encoder pairs and controlled acoustic conditions rather than a single model comparison. Our design therefore covers a broad set of encoders that preserve linguistic content while varying speaker identity and expressed emotion. We compare representations across layers and apply aggregation-aware permutation calibration to separate genuine cross-modal alignment from inflation of similarity introduced by searching over multiple layer pairs.

Local neighborhood agreement remains significant across all evaluated conditions after accounting for the search over layer pairs, and its magnitude is determined almost entirely by the speech encoder. In contrast, small-bandwidth RBF-CKA saturates on ℓ_2 -normalized representations and does not provide reliable evidence of shared global geometry. The strongest local alignment occurs between late speech layers and early text layers. Alignment does not increase consistently with model capability, is not associated with speaker or prosodic distance, and is not related to how well text encoders reproduce human semantic judgments. Taken together, these results indicate that the structure shared by speech and text encoders is tied to the sentence rather than to its acoustic realization, and that it corresponds more closely to lexical and surface-level representations than to fine-grained sentence meaning.

Contributions. Our contributions are threefold:

- We present, to our knowledge, the first aggregation-aware, layer-wise analysis of alignment between independently trained, unimodal speech and text encoders, without any shared objective or learned cross-modal mapping, covering 9 speech encoders, 10 text encoders, and all cross-modal layer pairs.
- We show that local neighborhood alignment survives permutation calibration across all evaluated conditions and is strongest between late speech layers and early text layers. We also identify the saturation of small-bandwidth RBF-CKA on normalized representations, which prevents it from providing reliable evidence about shared global geometry.
- We introduce a controlled speech-text corpus of

14,400 utterances selected from 57,600 candidates generated by four TTS systems and use it to examine how alignment relates to model capability, speaker identity, prosodic variation, and sentence-level semantic quality¹.

2 Related Work

The Platonic Representation Hypothesis (PRH) proposes that increasingly capable models trained on different modalities and objectives converge toward a shared statistical representation of reality (Huh et al., 2024). Its empirical support comes primarily from geometric similarities observed between the internal representations of vision and language models that were trained independently. Recent work, however, has challenged both the strength and interpretation of this evidence. Gröger et al. (2026) show that model width affects the similarity expected under random correspondence. They also examine the common practice of computing similarity across every pair of layers and reporting the highest value as the model-level score. Because this maximum tends to increase with the number of layer pairs searched, comparisons involving models of different depths can be artificially inflated. To account for both effects, they propose a permutation-based calibration that applies the same layer-selection procedure to the null distribution. After permutation-based calibration, evidence for global geometric convergence weakens substantially, whereas agreement between local neighborhoods remains significant, motivating the Aristotelian Representation Hypothesis (ARH). Koepke et al. (2026) further show that mutual-neighbor agreement decreases as the comparison set grows and under many-to-many correspondences, indicating that the remaining alignment is sensitive to the evaluation regime.

Taken together, these studies shift the focus from whether representations appear similar to whether the metrics and evaluation protocols used to establish that similarity provide reliable evidence of convergence. The original convergence analyses rely heavily on Centered Kernel Alignment (CKA) (Kornblith et al., 2019) and nearest-neighbor overlap (Huh et al., 2024). Ding et al. (2021) show that commonly used similarity measures can disagree on basic properties of learned representations

¹Code and corpus are available: <https://github.com/alefiury/Echoes-Platos-Cave> and <https://huggingface.co/datasets/alefiury/Echoes-Platos-Cave>

and propose evaluating them according to their sensitivity to functionally meaningful changes and specificity to irrelevant ones. [Davari et al. \(2022\)](#) further demonstrate that CKA is sensitive to outliers and dominant low-dimensional components and can change without corresponding changes in model behavior. Gaussian Radial Basis Function CKA (RBF-CKA) extends CKA with a nonlinear kernel that emphasizes relationships at a distance scale controlled by its bandwidth σ . [Alvarez \(2022\)](#) shows that RBF-CKA approaches linear CKA as σ becomes large. We complement this analysis by examining its behavior at small values of σ on ℓ_2 -normalized speech representations. Together, these results motivate distinguishing local from global alignment and accounting for the selection bias introduced by choosing the highest score across multiple layer pairs.

Layer-wise analyses show that neither speech nor text can be adequately characterized by final-layer representations alone. In text encoders, linguistic information is distributed across depth, with intermediate layers sometimes providing stronger representations than final layers ([Tenney et al., 2019](#); [Liu et al., 2019](#); [Voita et al., 2019](#); [Van Aken et al., 2019](#); [Skean et al., 2025](#)). Speech encoders similarly distribute acoustic, phonetic, lexical, and semantic information differently across layers ([Pasad et al., 2021, 2023, 2024](#)). They also retain speaker, prosodic, and paralinguistic information throughout the network: speaker and phone related structures exhibit varying degrees of separation ([Mohamed et al., 2024](#); [Chiu et al., 2025](#)), while suprasegmental properties such as lexical tone, stress, and phrasal accent are often strongest in intermediate layers ([de la Fuente and Jurafsky, 2024](#)). Because measured cross-modal alignment can depend on how consistently the same input is represented within each modality ([Hosseini et al., 2026](#)), we compare all speech-text layer pairs while holding linguistic content constant and varying speaker identity and expressed emotion.

Direct speech-text alignment has primarily been examined within models that couple the two modalities through shared training. Analyses of multi-modal foundation models and end-to-end speech language models find that speech and text representations become more similar across depth, while also identifying persistent gaps caused by temporal redundancy, language coverage, and late-layer decision instability ([Lee et al., 2024](#); [Hsu et al., 2026](#)). Text-only language models also contain au-

ditary information that can be recovered through a learned contrastive probe ([Ngo and Kim, 2024](#)). Other studies explicitly introduce speech-text alignment as a training objective for spoken-language understanding and speech generation ([Laperrière et al., 2023](#); [Choi et al., 2025](#); [Niu et al., 2025](#)). Recent results further show that the effectiveness of this supervision depends on the layer at which it is applied and that the layers storing the most information are not necessarily those where alignment is most beneficial ([Zhang et al., 2026](#); [Ye et al., 2026](#)). Unlike these settings, our study measures alignment between independently trained unimodal speech and text encoders without a shared objective or learned cross-modal mapping, while controlling for layer selection and acoustic variation.

3 Method

Our experimental design holds linguistic content constant while varying speaker identity and target emotion. This controlled setting allows us to examine whether speech-text alignment changes across acoustic variations of the same sentence while limiting variation caused by lexical and semantic content. The method consists of constructing a fully crossed synthetic speech corpus, extracting representations from every layer of multiple speech and text encoders, comparing these representations using complementary local and global metrics, and calibrating the strongest alignment observed across layer pairs.

3.1 Dataset Construction

We select 600 unique sentences from the SICK dataset ([Marelli et al., 2014](#)). SICK provides human semantic-relatedness judgments, which we later use to evaluate whether representational alignment is associated with semantic quality. The sentences are sampled from low-, medium-, and high-relatedness pairs while ensuring that no sentence occurs in more than one selected pair. The complete filtering and sampling procedure is described in Appendix A.1.

Speaker identity and target emotion are specified using reference recordings from RAVDESS ([Livingstone and Russo, 2018](#)). We randomly select three male and three female actors and retain four emotions: neutral, happy, sad, and angry. All reference recordings contain the same statement, ensuring that prompt content remains constant across speakers and emotions. Candidate prompts are

verified using a state-of-the-art (SOTA) speech emotion recognition (SER) model, emotion2vec+ Large (Ma et al., 2023). One recording is kept for each speaker-emotion combination, producing 24 prompt utterances.

Each sentence is synthesized under all 24 reference conditions using four SOTA zero-shot text-to-speech (TTS) models, namely, IndexTTS2 (Zhou et al., 2025), OmniVoice (Zhu et al., 2026), Qwen3-TTS (Hu et al., 2026), and VoxCPM2 (Zhou et al., 2026), producing a total of 57,600 utterances. Multiple systems are used because emotion-generation performance is model- and emotion-dependent, meaning that a TTS system that performs well at synthesizing or emulating one emotion may not be the strongest for another (Shen et al., 2026). Therefore, by using multiple TTS systems and filtering the most emotion-aligned utterances, we obtain a higher-quality, more emotion-aligned synthetic dataset for our controlled experiments.

For each (sentence, speaker, emotion) combination, we retain one candidate using a deterministic ranking that prioritizes agreement with the target emotion, followed by target-emotion confidence and average transcription accuracy. Emotion is evaluated using emotion2vec+ Large, while content preservation is measured using three SOTA Automatic Speech Recognition (ASR) models in order to compensate for possible transcription errors: Parakeet-TDT-0.6B-v3 (Sekoyan et al., 2025), Qwen3-ASR-1.7B (Shi et al., 2026), and Whisper-large-v3 (Radford et al., 2023). When no candidate is classified as the target emotion, the candidate with the highest score for the intended emotion is chosen. The resulting corpus contains 14,400 utterances. The corpus is available at <https://huggingface.co/datasets/alefiury/Echoes-Platos-Cave>. Complete per-system results are reported in Appendix A.4.

3.2 Models and Feature Extraction

We evaluate nine speech encoders: wav2vec 2.0 Base and Large (Baevski et al., 2020), HuBERT Large and X-Large (Hsu et al., 2021), WavLM Base+ and Large (Chen et al., 2022a), and Whisper Small, Medium, and Large-v3 (Radford et al., 2023). The ten text encoders are BERT Base and Large (Devlin et al., 2019), E5-v2 Small, Base, and Large (Wang et al., 2022), GTE-v1.5 Base and Large (Zhang et al., 2024), and Qwen3-Embedding 0.6B, 4B, and 8B (Zhang et al., 2025).

We extract every hidden layer, including the in-

put embedding output and front-end acoustic encoding. Speech representations are mean-pooled over time, and text representations over valid non-special tokens. The full design contains 2,160 conditions, defined by 9 speech encoders, 10 text encoders, 6 speakers, and 4 emotions. Each condition compares representations of the same 600 sentences across all speech-text layer pairs. Extraction and analysis code is available at <https://github.com/alefiury/Echoes-Platos-Cave>.

3.3 Representational Similarity

We evaluate both local neighborhood agreement and global geometric similarity. Local agreement asks whether the sentences that are nearest to a given sentence in one model are also its nearest neighbors in the other, regardless of how the two spaces are shaped overall. Global similarity asks whether the two spaces share the same overall geometry, so that all pairwise relations, including those between distant sentences, are preserved. Our primary local metric is mutual k -nearest-neighbor overlap:

$$\text{mKNN}_k(X_i, Y_j) = \frac{1}{Nk} \sum_{n=1}^N \left| \mathcal{N}_k^{X_i}(n) \cap \mathcal{N}_k^{Y_j}(n) \right|,$$

where $\mathcal{N}_k^{X_i}(n)$ and $\mathcal{N}_k^{Y_j}(n)$ are the k nearest neighbors of sentence n in speech layer i and text layer j . We evaluate $k \in \{10, 20, 50, 100\}$ and report both raw and chance-adjusted scores. We also compute cycle- k NN and CKNN at $k = 10$ (Huh et al., 2024).

Geometry-based comparisons use linear CKA, RSA with Spearman correlation, and RBF-CKA. RBF-CKA is evaluated after ℓ_2 normalization at $\sigma \in \{0.1, 0.5, 2, 5\}$ because its behavior depends on the kernel bandwidth and representation geometry (Alvarez, 2022). Complete metric definitions are provided in Appendix B.3.

3.4 Aggregation-Aware Calibration

Because the strongest alignment may occur at different depths, we compute each metric for every speech-text layer pair and use the maximum as the condition-level score. Since this maximum increases with the number of layer pairs searched, we apply the aggregation-aware permutation calibration of Gröger et al. (2026):

$$s_{\text{obs}} = \max_{i,j} m(X_i, Y_j),$$

$$s_{\text{null}}^{(r)} = \max_{i,j} m(X_i, P_{\pi_r} Y_j),$$

where P_{π_r} permutes the correspondence between speech and text samples. Applying the same layer-pair maximization to each permutation accounts for both random correspondence and selection over model depth.

We use 200 permutations and report the difference from the null mean, the margin above its 95th percentile, and permutation significance after Benjamini–Hochberg correction. Confirmatory calibration is applied to mutual k NN at $k = 10$ and RBF-CKA at $\sigma = 0.1$. Further details are provided in Appendix B.4.

3.5 Auxiliary Analyses

Speaker variation is measured using cosine distance between ECAPA2 embeddings (Thienpondt and Demuynck, 2023). Prosodic distance is computed from F0 and voicing features estimated with SwiftF0 (Nieradzik, 2025), together with intensity, speaking-rate, and duration features extracted with Praat (Boersma and Weenink, 2021). The features are standardized within speaker, and each emotional realization is compared with the corresponding neutral realization from the same speaker and TTS system. We also examine sample-size sensitivity, within-speech neighborhood stability, and the relationship between text-representation similarities and SICK human relatedness judgments.

Finally, we compare calibrated alignment with external capability measures from MTEB (Muenighoff et al., 2023) and SUPERB (Yang et al., 2021). Full definitions and parameter settings are reported in Appendix B.5.

4 Results and Discussion

We evaluate the strength, layer-wise organization, and model dependence of speech-text alignment. Additional results and robustness analyses are reported in Appendix C.

4.1 Local alignment survives calibration

Raw uncalibrated scores suggest agreement between speech and text representations across several neighborhood and geometry-based metrics. Table 1 summarizes the metrics most relevant to the main analysis, while complete results across all

Metric	Raw	Chance	Adj.	Cal.
<i>Local and neighborhood-based</i>				
Mutual k NN				
$k = 10$	0.249 ± 0.035	0.017	0.236 ± 0.035	0.231 ± 0.036
$k = 100$	0.371 ± 0.031	0.167	0.246 ± 0.037	—
CKNNA				
$k = 10$	0.295 ± 0.038	—	—	—
<i>Geometry-based</i>				
Linear CKA	0.398 ± 0.095	—	—	—
RSA				
Spearman	0.345 ± 0.064	—	—	—
RBF-CKA				
$\sigma = 0.5$	0.471 ± 0.071	—	—	—
$\sigma = 0.1$	0.986 ± 0.017	—	—	0.072 ± 0.056

Table 1: Representative maximum layer-pair scores averaged across the 90 speech-text encoder pairs, with each encoder-pair value summarizing 24 speaker-emotion conditions. Raw denotes the uncorrected maximum similarity across layer pairs. Chance is the expected overlap between random neighborhoods, $k/(N - 1)$, with $N = 600$. Adjusted denotes chance-corrected mutual k NN, computed as $(\text{Raw} - \text{Chance}) / (1 - \text{Chance})$. Calibrated denotes the aggregation-aware permutation score, which accounts for selecting the maximum across multiple layer pairs. Values are mean $\pm$ standard deviation; dashes indicate quantities that were not computed.

Quantity	Mutual k NN $k = 10$	RBF-CKA $\sigma = 0.1$
Critical threshold τ	$0.023 \pm <0.001$	0.984 ± 0.022
Margin over τ	0.226 ± 0.051	0.002 ± 0.004
Retention	$92.4 \pm 2.1\%$	$7.4 \pm 7.7\%$
Significant after BH-FDR	2,160/2,160	2,159/2,160

Table 2: Aggregation-aware calibration across 2,160 conditions. Mutual k NN remains well separated from its null threshold, whereas RBF-CKA shows only a small margin.

neighborhood sizes and RBF bandwidths are reported in Appendix C.1 (Table 8).

For mutual k NN at $k = 10$, the expected overlap between two random neighborhoods is $k/(N - 1) = 10/599 = 0.017$, whereas the observed score is approximately 0.25. Accounting for this baseline yields an adjusted score of 0.236. After aggregation-aware permutation calibration, the mean score remains 0.231, retaining 92.4% of its original magnitude. All 2,160 conditions remain significant after BH-FDR correction, and the mean margin between the observed score and its permutation threshold is 0.226 (Table 2). Local neighborhood agreement therefore remains strong after correcting for the tendency of the maximum score to increase when more layer pairs are compared.

RBF-CKA at $\sigma = 0.1$ behaves differently. Al-

though 2,159 conditions remain significant after BH-FDR correction, its mean calibrated score is only 0.072, retaining 7.4% of the raw value. Its mean margin over the permutation threshold is also only 0.002. Similar significance rates therefore conceal markedly different separation from the corresponding null distributions.

Figure 1 shows that the RBF-CKA result arises from kernel saturation. Its mean score increases abruptly from 0.471 at $\sigma = 0.5$ to 0.986 at $\sigma = 0.1$. At larger bandwidths, RBF-CKA increasingly approaches linear CKA, consistent with Alvarez (2022), and its ranking of encoder pairs becomes strongly correlated with the linear metric.

For ℓ_2 -normalized representations, a sufficiently small bandwidth drives most off-diagonal RBF kernel entries toward zero. The resulting kernel matrices approach identity-like structures and therefore appear highly aligned even when the underlying representations do not share meaningful geometry. The near-ceiling scores at $\sigma = 0.1$ should thus be treated as a measurement artifact rather than as evidence of strong cross-modal alignment. The calibrated analysis nevertheless provides robust evidence that independently trained speech and text encoders share local neighborhood structure.

4.2 Local alignment is highest in late speech layers and early text layers

Under mutual k NN at $k = 10$, the highest alignment scores are typically obtained by pairing late speech layers with early text layers. Figure 2 shows this pattern expressing layer positions as percentages of encoder depth.

The concentration in later speech layers is consistent with previous work on self-supervised layer-wise analysis of speech encoders, showing that acoustic, phonetic, and word-level information changes systematically across depth, while the layers at which different properties are most accessible depend on the architecture, model size, and pre-training objective (Pasad et al., 2021, 2023, 2024).

The concentration in early text layers is compatible with analyses showing that lower text-encoder layers retain comparatively more surface and phrase-level information, whereas later layers produce increasingly contextualized representations and encode more complex linguistic relations (Jawahar et al., 2019; Tenney et al., 2019; Ethayarajh, 2019). However, the organization of linguistic information across text layers does not follow a strict or universal progression, and

later analyses have questioned interpretations in which each depth corresponds to a distinct linguistic stage (Niu et al., 2022). Accordingly, the selection of early text layers does not identify a specific linguistic level shared with speech. It shows only that, for the neighborhood structure measured here, the closest correspondence with speech occurs before the final stages of text contextualization. Notably, for CKNN and small-bandwidth RBF-CKA the maximum most often falls on the text embedding output layer (Table 10). After pooling, this representation can be viewed as a bag-of-words.

4.3 Alignment is governed by the speech encoder

The speech encoder accounts for 95.4% of the variation among the 90 encoder-pair means under mutual k NN, while the text encoder accounts for only 3.3% (Table 11). Figure 3 shows the encoder-level scores before and after calibration. Speech encoders span a wide range, from 0.173 for wav2vec 2.0 Base to 0.283 for HuBERT X-Large, whereas text encoders fall within 0.219 to 0.242. Calibration lowers all scores by a similar amount and preserves both rankings (Appendix C.4). Which speech encoder is used therefore largely determines how well the modalities align, while the choice of text encoder matters comparatively little. Rankings do depend on the alignment metric, however, and the top encoder pairs under mutual k NN and linear CKA do not overlap (Appendix C.8).

4.4 Alignment does not increase consistently with model capability

A common prediction derived from the PRH is that more capable models should exhibit stronger representational alignment. The results do not show a consistent relationship across modalities. Among the five self-supervised speech encoders with comparable SUPERB evaluations, mutual k NN alignment increases with capability (Figure 4). SUPERB scores were taken from the official leaderboard and the original papers, the remaining speech encoders are omitted because no score was found in either source. However, the association is not statistically conclusive ($p > 0.05$), possibly because of the small number of samples ($\rho = 0.700$, $p = 0.19$).

The relationship is reversed for text encoders. Mutual k NN alignment decreases as MTEB performance increases across the ten evaluated models

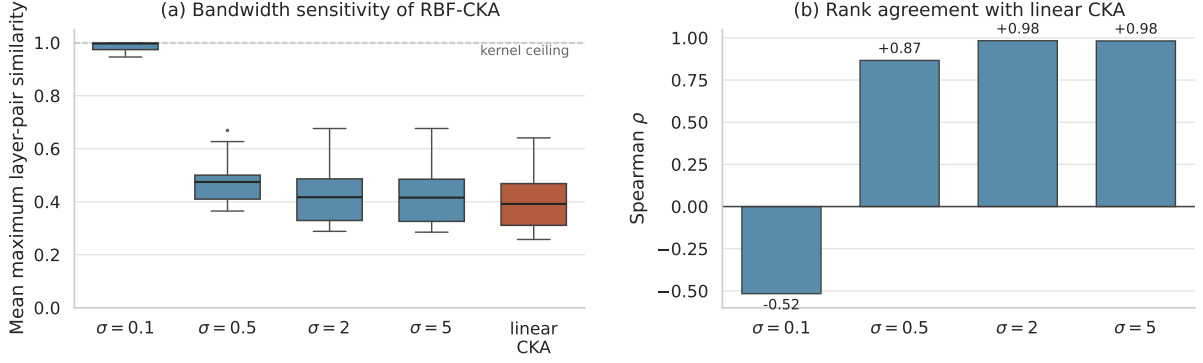


Figure 1: Sensitivity of RBF-CKA to kernel bandwidth. Small bandwidths produce near-ceiling scores and reverse the model ranking relative to linear CKA, whereas larger bandwidths recover strong rank agreement with the linear metric.

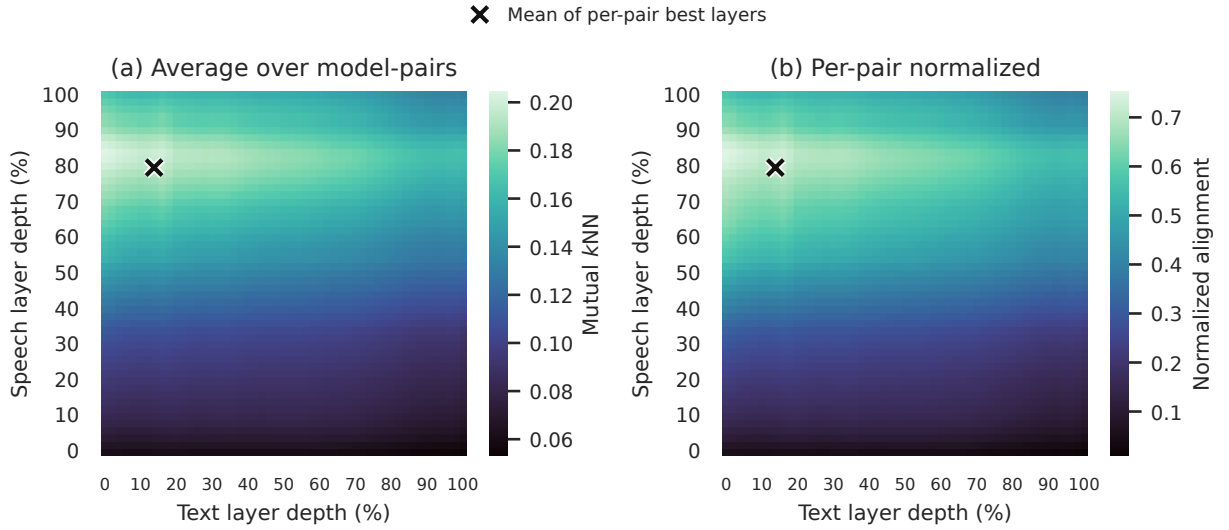


Figure 2: Depth-wise mutual k NN alignment across speech-text encoder pairs, averaged directly (a) or after normalizing scores within each pair (b). Crosses mark the mean position of the maximizing layer pairs. Both aggregations place the highest alignment in late speech layers and early text layers.

($\rho = -0.860$, $p = 0.007$). The association remains negative when the analysis is restricted to models trained explicitly for text embedding. It is also observed in the raw scores and when each speech encoder is analyzed separately, indicating that it is not introduced by the permutation calibration.

These results do not support a modality-independent, monotonic increase in speech-text alignment with model capability. Higher benchmark performance may therefore reflect changes in representational structure that do not necessarily increase cross-modal neighborhood agreement. In particular, objectives designed to improve text retrieval may organize local neighborhoods around distinctions that are useful for retrieval but are not preserved in independently trained speech representations.

4.5 Alignment magnitude is not associated with speaker or prosodic distance

Local alignment does not vary systematically with speaker distance. Across the 15 speaker pairs, ECAPA2 distance is not associated with mutual k NN similarity ($\rho = 0.139$, $p = 0.621$), and the corresponding Pearson correlation is approximately zero ($r = 0.001$, $p = 0.996$). Same-gender and cross-gender pairs also have similar mean scores (0.160 and 0.158, respectively), as reported in Appendix C.6. Additionally, this invariance to speaker identity and prosodic variation is not specific to self-supervised objectives. The supervised Whisper encoders show the same absence of association, which suggests that the shared structure recovered by speech encoders trained at scale is tied to sentence content regardless of the training objective.

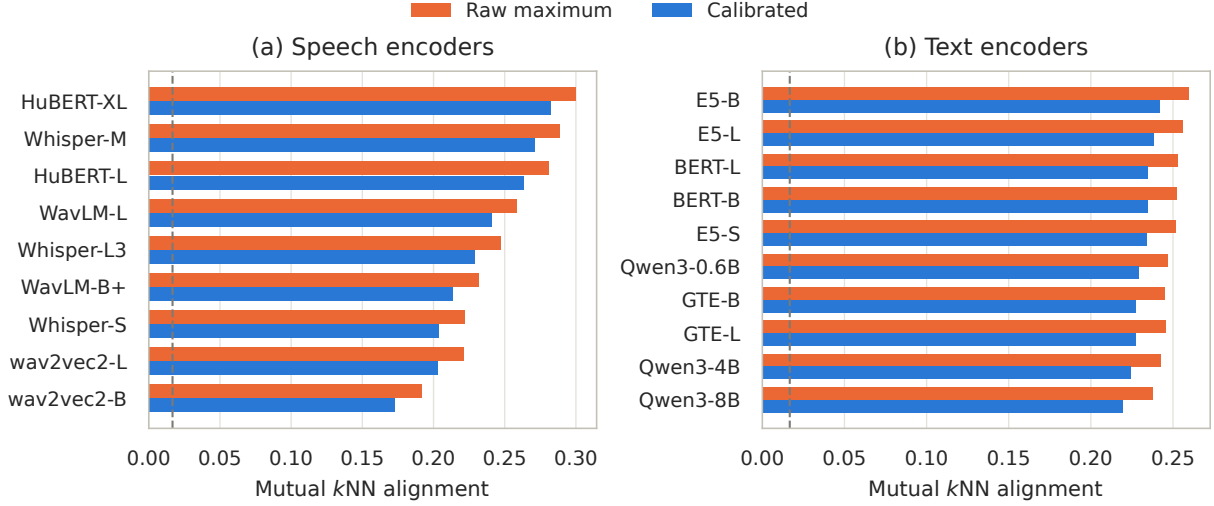


Figure 3: Raw and calibrated mutual k NN alignment by speech encoder (a) and text encoder (b). The dashed line marks the random-neighborhood baseline. Calibration preserves the rankings, and speech encoders span a much wider range than text encoders.

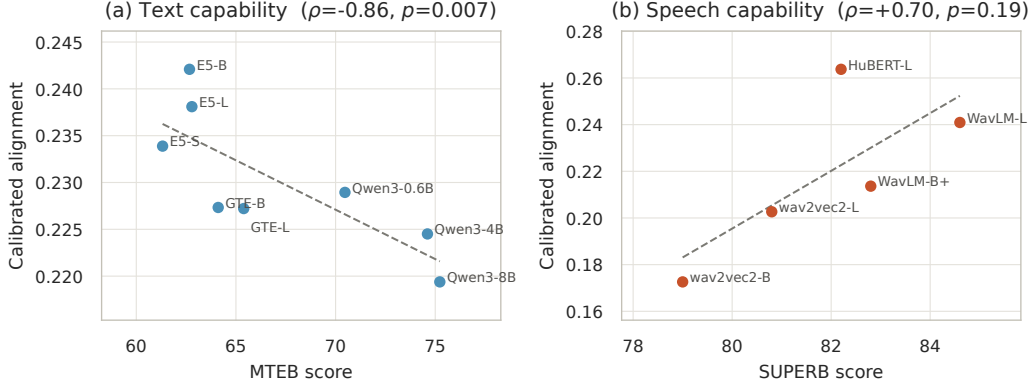


Figure 4: Calibrated mutual k NN alignment against model capability, using MTEB scores for text encoders (a) and SUPERB scores for self-supervised speech encoders (b), taken from the official leaderboards and original papers. Alignment decreases with text-encoder capability, whereas the speech-side association is positive but uncertain ($p > 0.05$) with only five models.

Prosodic distance is likewise unrelated to the magnitude of calibrated alignment. Across the 24 speaker-emotion conditions, its association with mutual k NN is small and uncertain ($\rho = -0.189, p = 0.376$). Neutral conditions have zero distance by construction; after excluding them, the association is approximately zero ($\rho = -0.044, p = 0.861$). Thus, utterances that differ more strongly from the corresponding neutral realization do not systematically exhibit weaker speech-text alignment.

Prosodic distance is nevertheless associated with the location of the maximum on the text side. Among non-neutral conditions, greater prosodic distance corresponds to earlier text layers ($\rho = -0.740, p < 0.001$), whereas selected speech-

layer depth shows no clear relationship ($\rho = 0.127, p = 0.616$). This exploratory result suggests that acoustic realization may affect which text representations align most strongly with speech without substantially changing the magnitude of that alignment. Complete speaker- and prosody-level results are reported in Appendix C.6.

4.6 Alignment is not explained by semantic quality

Text encoders that better reproduce human judgments of sentence meaning do not align more strongly with speech. At their best layers, correlations with the SICK relatedness annotations range from 0.744 to 0.891. However, this semantic performance is not associated with calibrated mutual k NN alignment across models ($\rho = 0.10$,

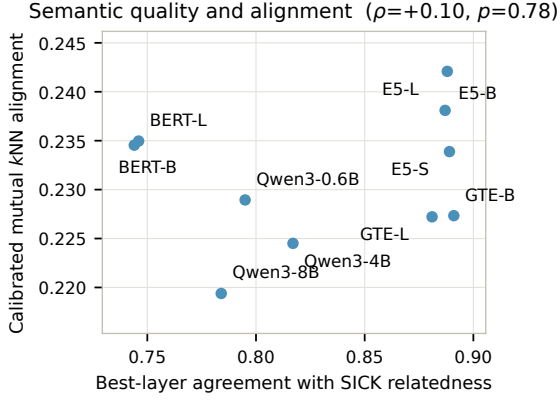


Figure 5: Relationship between sentence-level semantic quality and speech-text alignment across the ten text encoders. Semantic quality is the highest Spearman correlation between each encoder’s sentence similarities and the SICK human relatedness judgments. Alignment is the mean calibrated mutual k NN score at $k = 10$. Each point represents one text encoder. Semantic quality is not associated with speech-text alignment ($\rho = 0.10$, $p = 0.78$).

$p = 0.78$; Figure 5).

Although semantic performance varies substantially across text encoders, these differences are not associated with how strongly the encoders align with speech. Together with the frequent selection of early text layers, this result suggests that the observed agreement may partly reflect lexical, distributional, or broad categorical structure rather than fine-grained sentence meaning. Additional semantic and gallery-size analyses are reported in Appendix C.7.

4.7 Implications

Our results suggest where speech and text representations can be most effectively connected. The strongest local correspondence appears between late speech layers and early text layers, and its magnitude depends almost entirely on the speech encoder, so methods that bridge the two modalities, such as adapters for speech language models or alignment objectives for synthesis, can use these depths as a starting point and treat the choice of speech encoder as the main lever for improving compatibility. Since local agreement remains after calibration while global similarity does not, methods that rely on local structure may be more appropriate than those that assume the two spaces can be aligned globally. For example, retrieval-based approaches or nearest-neighbor mappings may be better suited than a single linear transformation.

The pattern of correspondence also gives some indication of what information is shared between the two modalities. Correspondence is strongest in the least contextualized text layers, is not related to how well a text encoder captures sentence meaning, and does not change systematically with acoustic variation. Together, these results suggest that the shared structure is mainly lexical rather than semantic or syntactic. This does not mean that speech encoders contain no semantic information. Instead, it suggests that the information they most clearly share with text encoders is primarily lexical. For future speech foundation models, stronger alignment with text beyond the word level may therefore require pretraining objectives that place greater emphasis on sentence-level semantic and syntactic information, rather than mainly capturing lexical and acoustic content.

5 Conclusion

We examined representational alignment between 9 speech encoders and 10 text encoders by comparing all speech-text layer pairs across 2,160 controlled conditions and calibrating the maximum similarity with an aggregation-aware permutation procedure. Local neighborhood agreement remained significant in every condition, was highest between late speech layers and early text layers, and was determined almost entirely by the speech encoder. By contrast, small-bandwidth RBF-CKA saturated on ℓ_2 -normalized representations and did not provide reliable evidence about shared global geometry. Alignment did not increase consistently with model capability, was not associated with speaker or prosodic distance, and was not related to how well text encoders reproduce human semantic judgments. Together, these results indicate that independently trained speech and text encoders share a robust local structure that is tied to the sentence rather than to its acoustic realization and that is primarily lexical rather than semantic, but they do not establish convergence toward a common global representation space. This shared structure offers a starting point for methods that connect the two modalities, and it suggests that alignment beyond the word level will require speech representations that capture more of sentence-level meaning.

6 Limitations

This study uses synthetic speech from six speakers, four emotions, and four TTS systems, so the results

may not transfer directly to natural speech or to more complex acoustic conditions. Target-emotion agreement is substantially lower for happy and sad speech than for angry and neutral speech, and synthesis provenance is not balanced across emotions, so the speaker and prosody analyses should be read as descriptive. Moreover, the SER and ASR models used to evaluate corpus quality also contribute to candidate selection, meaning that the reported quality scores are not fully independent.

The sentences come from a single English dataset whose pairs are built through minimal edits, so sentences that share words also tend to share meaning. Our data therefore cannot fully separate lexical overlap from semantic relatedness, and the lexical interpretation of the shared structure should be read as the simplest explanation consistent with the results rather than as a demonstrated one. Evaluation on other datasets and languages would test whether the pattern generalizes.

The permutation analysis uses 200 permutations, and only mutual k NN at $k = 10$ and RBF-CKA at $\sigma = 0.1$ undergo complete calibration. Because the latter is affected by kernel saturation, calibrated analyses of linear CKA, RSA, and other geometry-based measures are required before drawing stronger conclusions about global alignment.

Ethical Considerations

This work uses two publicly available research datasets, SICK (Marelli et al., 2014) and RAVDESS (Livingstone and Russo, 2018), as well as publicly released pretrained models, all used in accordance with their respective licenses. No new human data were collected, and no human annotators were employed. Emotion labels and transcription accuracy were assessed using automatic models rather than human raters. The implications of these choices are discussed in Section 6.

The controlled speech corpus was generated through zero-shot voice cloning of six RAVDESS actors. The original recordings were collected with informed consent for research purposes and are distributed under the CC BY-NC-SA 4.0 license. The cloned voices are used exclusively as controlled acoustic conditions for representational analysis. We do not attribute the synthesized utterances to the original speakers, nor should the generated content be interpreted as reflecting their views, beliefs, or intentions. The released corpus retains the RAVDESS license and its restrictions on com-

mercial use. Given the potential misuse of voice-cloning data, we further discourage its use for developing or evaluating systems intended for speaker impersonation or deception.

Pretrained models were used exclusively for research purposes and were neither fine-tuned nor redistributed. The study is restricted to English, and its findings should not be assumed to generalize to other languages without further evaluation. More broadly, our analyses characterize representations learned by existing models and do not introduce new capabilities for surveillance, deception, or profiling. To facilitate transparency and reproducibility, we release the experimental code, controlled corpus, and extracted representations used in the study.

Acknowledgments

This work has been funded by the project Research and Development of Genese Digital: Scaling Interactive and Culturally Adapted Digital Humans with Generative AI, supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI, grant number 057/2023, signed with EM-BRAPII. The authors gratefully acknowledge the Artificial Intelligence Lab at Recod.ai and the Institute of Computing at the University of Campinas for providing access to the computational infrastructure used in the experiments. The authors also acknowledge the support and contributions of Huglabs and Ermis.ai.

References

- Sergio A Alvarez. 2022. Gaussian RBF centered kernel alignment (CKA) in the large-bandwidth limit. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):6587–6593.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.
- Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 <http://www.praat.org/>.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma

- Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. [WavLM: Large-scale self-supervised pre-training for full-stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022b. BEATs: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.
- Aemon Yat Fei Chiu, Kei Ching Fung, Roger Tsz Yeung Li, Jingyu Li, and Tan Lee. 2025. A large-scale probing analysis of speaker-specific attributes in self-supervised speech representations. *arXiv preprint arXiv:2501.05310*.
- Jeongsoo Choi, Zhikang Niu, Ji-Hoon Kim, Chunhui Wang, Joon Son Chung, and Xie Chen. 2025. [Accelerating diffusion-based text-to-speech model training with dual modality alignment](#). *arXiv preprint arXiv:2505.19595*.
- MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2022. Reliability of CKA as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*.
- Antón de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models. *arXiv preprint arXiv:2408.13678*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. 2021. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Manat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.
- Fabian Gröger, Shuo Wen, and Maria Brbic. 2026. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*.
- Eghbal A Hosseini, Brian Cheung, Evelina Fedorenko, and Alex H Williams. 2026. Modulating cross-modal convergence with single-stimulus, intra-modal dispersion. *arXiv preprint arXiv:2604.21836*.
- Ming-Hao Hsu, Xueyao Zhang, Xiaohai Tian, Jun Zhang, and Zhizheng Wu. 2026. [Anatomy of the modality gap: Dissecting the internal states of end-to-end speech LLMs](#). *arXiv preprint arXiv:2603.01502*.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Hangrui Hu, Xinfu Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-TTS technical report. *arXiv preprint arXiv:2601.15621*.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. [What does BERT learn about the structure of language?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- A. Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. 2026. Back into plato’s cave: examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. [Similarity of neural network representations revisited](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Gaëlle Laperrière, Valentin Pelloin, Mickaël Rouvier, Themis Stafylakis, and Yannick Esteve. 2023. On the use of semantically-aligned speech representations for spoken language understanding. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 361–368. IEEE.

- Hyunji Lee, Danni Liu, Supriti Sinhamahapatra, and Jan Niehues. 2024. [How do multimodal foundation models encode text and speech? an analysis of cross-lingual and cross-modal representations](#). *arXiv preprint arXiv:2411.17666*.
- Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094.
- Steven R. Livingstone and Frank A. Russo. 2018. [The ryerson audio-visual database of emotional speech and song \(RAVDESS\): A dynamic, multimodal set of facial and vocal expressions in north american english](#). *PLOS ONE*, 13(5):e0196391.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. [A SICK cure for the evaluation of compositional distributional semantic models](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Mukhtar Mohamed, Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2024. Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations. *arXiv preprint arXiv:2406.09200*.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.
- Jerry Ngo and Yoon Kim. 2024. What do language models hear? probing for auditory representations in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5435–5448.
- Lars Nieradzki. 2025. [SwiftF0: Fast and accurate monophonic pitch detection](#). *arXiv preprint arXiv:2508.18440*.
- Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. [Does BERT rediscover a classical NLP pipeline?](#) In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Zhikang Niu, Shujie Hu, Jeongsoo Choi, Yushen Chen, Peining Chen, Pengcheng Zhu, Yunting Yang, Bowen Zhang, Jian Zhao, Chunhui Wang, and Xie Chen. 2025. [Semantic-VAE: Semantic-alignment latent representation for better speech synthesis](#). *arXiv preprint arXiv:2509.22167*.
- Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. [What do self-supervised speech models know about words?](#) *Transactions of the Association for Computational Linguistics*, 12:372–391.
- Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.
- Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. [Canary-1B-v2 & Parakeet-TDT-0.6B-v3: Efficient and high-performance models for multilingual ASR and AST](#). *arXiv preprint arXiv:2509.14128*.
- Shengfan Shen, Di Wu, Xingchen Song, Dinghao Zhou, Liumeng Xue, Meng Meng, Jian Luan, and Shuai Wang. 2026. Iterate to differentiate: Enhancing discriminability and reliability in zero-shot tts evaluation. *arXiv preprint arXiv:2603.24430*.
- Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*.
- Oscar SKEAN, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*.

- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.
- Jenthe Thienpondt and Kris Demuynck. 2023. ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1823–1832.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Kotik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. **SUPERB: Speech processing universal performance benchmark**. *arXiv preprint arXiv:2105.01051*.
- Zhen Ye, Xu Tan, Yiming Li, Guangyan Zhang, Chimin Chan, Haohe Liu, Zhengxi Liu, Hongzhan Lin, Zheqi Dai, Xinshen Zhang, Peiwen Sun, Qiuqiang Kong, and Wei Xue. 2026. **Which speech representation better matches text-native reasoning? a study of speech-text alignment on frame rate and representation**. *arXiv preprint arXiv:2606.12199*.
- Pengfei Zhang, Tianxin Xie, Minghao Yang, and Li Liu. 2026. AG-REPA: Causal layer selection for representation alignment in audio flow matching. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. **IndexTTS2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech**. *arXiv preprint arXiv:2506.21619*.
- Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, Zhisheng Zhang, Jiuyang Zhou, Bingsong Bai, Weiye Sun, Mengyuan Deng, Qundong Shi, Zhiyong Wu, and Zhiyuan Liu. 2026. VoxCPM2 technical report. *arXiv preprint arXiv:2606.06928*.
- Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. 2026. OmniVoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*.

A Corpus Construction and Quality

A.1 Sentence Selection

We combine the training, validation, and test splits of SICK and remove pairs in which either sentence contains fewer than 4 or more than 30 words. We also remove identical pairs, reversed duplicates, and sentences containing questions, exclamations, URLs, or HTML content.

The remaining pairs are divided into three semantic-relatedness bands: low (≤ 2.3), medium (3.0–3.7), and high (≥ 4.2). The transition intervals (2.3, 3.0) and (3.7, 4.2) are excluded to maintain separation between adjacent bands. We sample 100 pairs from each band while requiring each sentence to occur in only one selected pair. Disjoint sampling is attempted up to 1,000 times.

We do not rebalance the entailment labels because entailment and semantic relatedness are not independent in SICK. The resulting set contains 300 pairs and 600 unique sentences. Mean relatedness is 1.44, 3.34, and 4.65 in the low, medium, and high bands, respectively.

A.2 Reference-Prompt Selection

We randomly select three male and three female RAVDESS actors and retain the neutral, happy, sad, and angry conditions, which are available for all six actors. All reference recordings contain the statement “Dogs are sitting by the door,” so the

linguistic content of the prompt remains constant across speakers and emotions.

Each candidate reference is evaluated using `emotion2vec+ Large`². Recordings for which the predicted emotion disagrees with the RAVDESS annotation are excluded. Among the remaining candidates, one recording is retained for each speaker-emotion combination by prioritizing emotional intensity and then the confidence assigned to the annotated emotion. This procedure produces 24 reference recordings.

A.3 Speech Synthesis and Candidate Selection

Every sentence is synthesized for each of the 24 speaker-emotion reference conditions using IndexTTS2, OmniVoice, Qwen3-TTS, and VoxCPM2. Each system receives the same reference waveform. Systems that support reference-text conditioning additionally receive the transcript of the RAVDESS prompt. The complete generation procedure produces $600 \times 24 \times 4 = 57,600$ candidate utterances.

For each (sentence, speaker, emotion) combination, the four candidates are ranked. The ranking criteria are, in order:

1. Agreement between the intended emotion and the prediction of `emotion2vec+ Large`;
2. Confidence assigned to the intended emotion;
3. Mean WER across Parakeet-TDT-0.6B-v3, Qwen3-ASR-1.7B, and Whisper-large-v3;
4. Mean CER across the same ASR systems.

Higher emotion confidence and lower error rates are preferred. A fixed ordering of the synthesis systems is used only when all preceding criteria are tied in the following order: OmniVoice, VoxCPM2, IndexTTS2, and Qwen3-TTS. The final filtered corpus contains 14,400 utterances.

A.4 Corpus Quality

Table 3 reports agreement between the intended emotion and the prediction of `emotion2vec+ Large`. Per-condition selection increases mean target-emotion agreement from 67.500% for the strongest individual system to 80.618%.

Table 4 reports transcription quality across the three ASR systems. All synthesis systems have median WER and CER equal to zero. In the selected

corpus, 92.611% of utterances are transcribed without word errors by all three systems, and 97.424% have mean WER below 10%.

A.5 Synthesis Provenance

OmniVoice contributes 5,960 selected utterances, VoxCPM2 contributes 4,403, IndexTTS2 contributes 2,109, and Qwen3-TTS contributes 1,928. Table 5 shows that this distribution differs across intended emotions. Synthesis provenance is therefore not independently balanced across the speaker-emotion conditions and should not be interpreted as a controlled experimental factor.

B Models and Analysis Details

B.1 Model Checkpoints

Tables 6 and 7 list the evaluated checkpoints and the number of representations extracted from each model. The counts include the input embedding output or front-end acoustic output as layer 0, followed by the outputs of the main encoder or transformer blocks.

B.2 Feature Extraction

Audio is processed using the sampling rate and preprocessing procedure expected by each speech checkpoint. Hidden states are extracted from every available speech layer and mean-pooled over valid time steps. Text is tokenized using the corresponding tokenizer, and hidden states are mean-pooled over valid nonspecial tokens.

Each condition contains one representation for each of the same 600 sentences in speech and text. Nearest neighbors are computed using cosine similarity. Layer indices refer to the layer from which the representations were extracted.

B.3 Metric Settings

Mutual k NN is evaluated at $k \in \{10, 20, 50, 100\}$. Nearest-neighbor sets exclude the sample itself. The overlap expected between two random neighborhoods is

$$c_k = \frac{k}{N - 1}.$$

For $N = 600$, this equals 0.0167, 0.0334, 0.0835, and 0.1669 for the four values of k . The chance-adjusted score is

$$\text{mKNN}_{\text{adj}} = \frac{\text{mKNN} - c_k}{1 - c_k}.$$

²https://huggingface.co/emotion2vec/emotion2vec_plus_large

System	Angry	Happy	Neutral	Sad	Mean
IndexTTS2	90.889 $\pm$ 10.040	35.806 $\pm$ 22.311	92.528 $\pm$ 4.819	28.583 $\pm$ 35.031	61.951 $\pm$ 15.312
OmniVoice	92.861 $\pm$ 6.739	38.000 $\pm$ 21.780	96.861 $\pm$ 3.735	40.611 $\pm$ 33.757	67.083 $\pm$ 14.383
Qwen3-TTS	56.722 $\pm$ 14.349	27.278 $\pm$ 21.155	94.889 $\pm$ 3.423	20.028 $\pm$ 13.588	49.729 $\pm$ 9.140
VoxCPM2	96.194 $\pm$ 3.938	33.500 $\pm$ 21.436	94.111 $\pm$ 10.707	46.194 $\pm$ 38.354	67.500 $\pm$ 16.581
Mean of systems	84.167 $\pm$ 5.851	33.646 $\pm$ 17.530	94.597 $\pm$ 5.253	33.854 $\pm$ 28.911	61.566 $\pm$ 12.471
Selected corpus	99.056 $\pm$ 1.237	63.750 $\pm$ 22.413	99.361 $\pm$ 0.710	60.306 $\pm$ 40.544	80.618 $\pm$ 15.201

Table 3: Agreement (%) between the intended emotion and the prediction of emotion2vec+ Large. Values are mean $\pm$ standard deviation across the six speakers, with each speaker-level score computed over 600 sentences per emotion. Mean is the macro-average across the four emotions. Mean of systems pools the four synthesis systems within each speaker before aggregation. Selected corpus denotes the utterances retained after per-condition candidate selection. Because the same classifier contributes to candidate selection, these values measure agreement with the selection criterion rather than an independent perceptual evaluation.

System	WER (%)		CER (%)		Exact (%)
	Mean	Median	Mean	Median	
IndexTTS2	1.538 $\pm$ 2.030	0.000 $\pm$ 0.000	0.651 $\pm$ 1.209	0.000 $\pm$ 0.000	91.139 $\pm$ 4.069
OmniVoice	0.607 $\pm$ 0.103	0.000 $\pm$ 0.000	0.125 $\pm$ 0.038	0.000 $\pm$ 0.000	93.979 $\pm$ 0.833
Qwen3-TTS	0.821 $\pm$ 0.091	0.000 $\pm$ 0.000	0.261 $\pm$ 0.085	0.000 $\pm$ 0.000	90.722 $\pm$ 1.662
VoxCPM2	0.929 $\pm$ 0.279	0.000 $\pm$ 0.000	0.363 $\pm$ 0.286	0.000 $\pm$ 0.000	91.903 $\pm$ 0.967
Selected corpus	0.929 $\pm$ 0.673	0.000 $\pm$ 0.000	0.336 $\pm$ 0.429	0.000 $\pm$ 0.000	92.611 $\pm$ 1.661

Table 4: Transcription quality measured using Parakeet-TDT-0.6B-v3, Qwen3-ASR-1.7B, and Whisper-large-v3. For each speaker, utterance-level WER and CER are first averaged across the three ASR systems and then summarized across utterances. Table entries are the mean $\pm$ standard deviation of these summaries across the six speakers. Exact is the percentage of utterances transcribed without word errors by all three systems.

This transformation maps the expected random overlap to zero while retaining one as the maximum possible value.

Cycle k NN and unbiased CKNNA are evaluated at $k = 10$. Linear CKA is computed from centered linear Gram matrices. RSA is computed as the Spearman correlation between the upper-triangular entries of cosine-distance matrices. RBF-CKA is computed after ℓ_2 normalization using

$$K_{ab} = \exp\left(-\frac{\|z_a - z_b\|_2^2}{2\sigma^2}\right),$$

with $\sigma \in \{0.1, 0.5, 2, 5\}$.

The maximum over all speech-text layer pairs is used as the condition-level aggregate. Depending on model depth, a condition compares between $13 \times 13 = 169$ and $49 \times 37 = 1,813$ layer pairs.

B.4 Calibration Details

For each condition and calibrated metric, the observed aggregate is

$$s_{\text{obs}} = \max_{i,j} m(X_i, Y_j).$$

The correspondence between speech and text samples is then randomly permuted. The same

permutation is applied across all layers of the text model, and the maximum is recomputed:

$$s_{\text{null}}^{(r)} = \max_{i,j} m(X_i, P_{\pi_r} Y_j).$$

Applying the same maximization to the observed and permuted data accounts for the increase in the maximum expected when more layer pairs are searched.

We use 200 independent permutations. The critical threshold τ is the 95th percentile of the permutation distribution, and the threshold margin is

$$\Delta_\tau = s_{\text{obs}} - \tau.$$

Permutation p -values use the finite-sample correction

$$p = \frac{1 + \sum_{r=1}^R \mathbb{I}[s_{\text{null}}^{(r)} \geq s_{\text{obs}}]}{R + 1},$$

giving a minimum attainable value of $1/201 \approx 0.005$. Benjamini–Hochberg correction is applied separately to mutual k NN at $k = 10$ and RBF-CKA at $\sigma = 0.1$, with 2,160 tests within each metric. Seed 42 is used for all randomized procedures.

Emotion	IndexTTS2	OmniVoice	Qwen3-TTS	VoxCPM2
Angry	11.6	51.3	5.3	31.8
Happy	24.4	29.1	23.6	23.0
Neutral	8.5	48.4	10.3	32.8
Sad	14.1	36.7	14.4	34.8

Table 5: Percentage of selected utterances contributed by each synthesis system within each intended emotion.

Model	Checkpoint	Family	Num. Layers
wav2vec 2.0 Base	facebook/wav2vec2-base	wav2vec 2.0	13
wav2vec 2.0 Large	facebook/wav2vec2-large	wav2vec 2.0	25
HuBERT Large	facebook/hubert-large-1160k	HuBERT	25
HuBERT X-Large	facebook/hubert-xlarge-1160k	HuBERT	49
WavLM Base+	microsoft/wavlm-base-plus	WavLM	13
WavLM Large	microsoft/wavlm-large	WavLM	25
Whisper Small	openai/whisper-small	Whisper	13
Whisper Medium	openai/whisper-medium	Whisper	25
Whisper Large-v3	openai/whisper-large-v3	Whisper	33

Table 6: Speech checkpoints and numbers of extracted representations as presented on Hugging Face.

B.5 Auxiliary Analyses

Speaker distance is measured as cosine distance between speaker representations produced by the external ECAPA2 speaker encoder. Pairwise analyses contain the 15 unordered pairs formed by the six speakers. Because each speaker occurs in several pairs, these observations are not statistically independent; the corresponding correlations and group comparisons are therefore treated as descriptive.

Prosodic distance is computed from 15 features: F0 mean, F0 median, F0 standard deviation in hertz, F0 standard deviation in semitones, F0 range in semitones, the rate of F0 change, voiced fraction, voiced-segment rate, three F0-confidence statistics, intensity mean, intensity standard deviation, syllable rate, and duration. Each feature is standardized across the available speaker-emotion-system observations. For a speaker s , emotion e , and synthesis system t , prosodic distance is the Euclidean distance from the corresponding neutral vector:

$$d_{\text{pros}}(s, e, t) = \|\mathbf{z}_{s,e,t} - \mathbf{z}_{s,\text{neutral},t}\|_2.$$

Consequently, neutral conditions have $d_{\text{pros}} = 0$ by construction. Analyses are therefore reported both with and without neutral conditions. One (speaker, emotion, system) combination is unavailable, yielding 95 rather than 96 system-specific observations and 71 non-neutral observations.

Sample-size sensitivity is evaluated at $N \in \{128, 256, 512, 600\}$ using 10 random draws for $N < 600$. We compare a fixed neighborhood size

of $k = 10$ with $k = \text{round}(N/100)$. Both raw and chance-adjusted mutual k NN are reported.

For semantic evaluation, cosine similarities produced by each text layer are correlated with the human relatedness judgments for the 300 selected SICK pairs. The layer with the highest Spearman correlation is retained for each model. These semantic scores are then compared with model-level speech-text alignment.

C Additional Results

C.1 Complete Metric Results

Table 8 reports maximum layer-pair scores averaged across the 90 speech-text encoder pairs. Each encoder-pair value first averages the 24 speaker-emotion conditions.

C.2 Calibration Diagnostics

Table 9 reports condition-level calibration statistics. Mutual k NN remains substantially above its permutation threshold, whereas RBF-CKA at $\sigma = 0.1$ exceeds its threshold by only a small margin.

C.3 Layer-Wise Results

Table 10 summarizes the locations of the maximum raw layer-pair scores. Layer 0 denotes the input embedding output. Mutual k NN, CKNN, and small-bandwidth RBF-CKA frequently select the text embedding output or one of the first transformer representations.

Model	Checkpoint	Family	Num. Layers
BERT Base	google-bert/bert-base-uncased	BERT	13
BERT Large	google-bert/bert-large-uncased	BERT	25
E5-v2 Small	intfloat/e5-small-v2	E5	13
E5-v2 Base	intfloat/e5-base-v2	E5	13
E5-v2 Large	intfloat/e5-large-v2	E5	25
GTE-v1.5 Base	Alibaba-NLP/gte-base-en-v1.5	GTE	13
GTE-v1.5 Large	Alibaba-NLP/gte-large-en-v1.5	GTE	25
Qwen3-Embedding 0.6B	Qwen/Qwen3-Embedding-0.6B	Qwen3-Embedding	29
Qwen3-Embedding 4B	Qwen/Qwen3-Embedding-4B	Qwen3-Embedding	37
Qwen3-Embedding 8B	Qwen/Qwen3-Embedding-8B	Qwen3-Embedding	37

Table 7: Text checkpoints and numbers of extracted representations as presented on Hugging Face.

Metric	Raw score	Adjusted score	Calibrated score
<i>Local and neighborhood-based metrics</i>			
Mutual k NN			
$k = 10$	0.249 ± 0.035	0.236 ± 0.035	0.231 ± 0.036
$k = 20$	0.260 ± 0.035	0.235 ± 0.036	—
$k = 50$	0.304 ± 0.033	0.240 ± 0.036	—
$k = 100$	0.371 ± 0.031	0.246 ± 0.037	—
CKNNA			
$k = 10$	0.295 ± 0.038	—	—
Cycle k NN			
$k = 10$	0.814 ± 0.035	—	—
<i>Geometry-based metrics</i>			
Linear CKA	0.398 ± 0.095	—	—
RSA			
Spearman	0.345 ± 0.064	—	—
RBF-CKA			
$\sigma = 0.1$	0.986 ± 0.017	—	0.072 ± 0.056
$\sigma = 0.5$	0.471 ± 0.071	—	—
$\sigma = 2$	0.425 ± 0.100	—	—
$\sigma = 5$	0.423 ± 0.101	—	—

Table 8: Maximum layer-pair scores across the 90 speech-text encoder pairs, with each pair summarizing 24 speaker-emotion conditions. Raw is the uncorrected maximum. Adjusted is the chance-corrected mutual k NN score. Calibrated accounts for selecting the maximum across multiple layer pairs using an aggregation-aware permutation null. Values are mean $\pm$ standard deviation across encoder pairs; dashes indicate quantities that were not computed.

C.4 Encoder Variation and Rankings

The decomposition in Table 11 is applied to the 9×10 matrix of raw mutual k NN encoder-pair means. It is descriptive and does not constitute a causal variance analysis.

Tables 12 and 13 report the encoder-level means before and after calibration. For speech encoders, each row summarizes the 10 text-encoder means. For text encoders, each row summarizes the 9 speech-encoder means.

C.5 Capability Analysis

Table 14 summarizes the benchmark-based capability analysis reported in Section 4.4. Spearman correlation is used because the external evaluations

provide ordered capability comparisons and may differ in scale across benchmark versions.

The negative text-side association is also present in the uncalibrated scores and when text encoders are compared separately within individual speech encoders. The speech-side estimate is based on only five models and is therefore reported as directional rather than conclusive.

C.6 Speaker and Prosodic Variation

Table 15 reports associations between ECAPA2 distance and three representation-similarity measures across the 15 speaker pairs. None shows a consistent relationship with speaker distance. Same-gender and cross-gender pairs also have similar

Quantity	Mutual k NN $k = 10$	RBF-CKA $\sigma = 0.1$
Observed score	0.249 ± 0.051	0.986 ± 0.019
Calibrated score	0.231 ± 0.053	0.072 ± 0.075
Critical threshold τ	$0.023 \pm <0.001$	0.984 ± 0.022
Margin over τ	0.226 ± 0.051	0.002 ± 0.004
Retention	$92.4 \pm 2.1\%$	$7.4 \pm 7.7\%$
Significant after BH-FDR	2,160/2,160	2,159/2,160

Table 9: Aggregation-aware calibration across 2,160 conditions. Values are mean $\pm$ standard deviation across conditions. Retention is the calibrated score divided by the observed score.

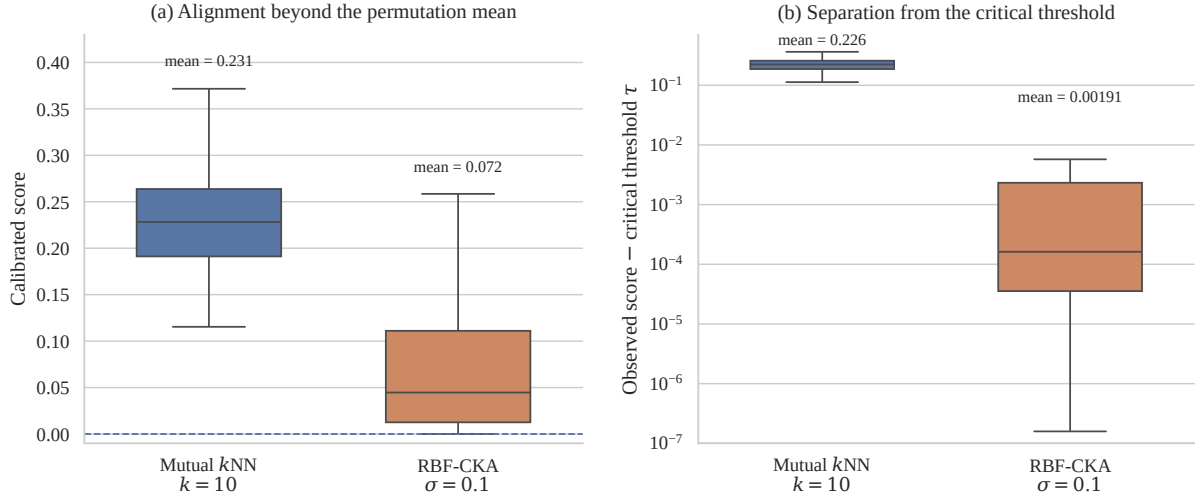


Figure 6: Calibration of local neighborhood agreement and RBF-CKA. Panel (a) shows calibrated scores across the 2,160 conditions. The dashed line indicates the random-neighborhood baseline for mutual k NN at $k = 10$. Panel (b) shows the margins between the observed scores and their permutation thresholds. Mutual k NN remains well separated from its null distribution, whereas RBF-CKA at $\sigma = 0.1$ exceeds its threshold by only a small margin.

mean mutual k NN scores: 0.160 ± 0.018 and 0.158 ± 0.020 , respectively. Their difference is not significant under a Mann–Whitney comparison ($p = 0.864$), although this test does not account for the reuse of speakers across pairs.

Prosodic distance is not associated with the magnitude of calibrated mutual k NN alignment (Table 16). This remains true after neutral conditions, whose distance is zero by construction, are excluded. Among non-neutral conditions, greater prosodic distance is associated with earlier text-layer maxima but not with speech-layer depth.

Table 17 summarizes the prosodic distance, calibrated alignment, and selected relative depths for each intended emotion. Angry speech is the most distant from neutral on average but does not have the lowest alignment, providing no evidence of a monotonic reduction in alignment as prosodic deviation increases.

Table 18 identifies the features most strongly associated with the prosodic-distance summary.

These correlations are descriptive because each feature contributes directly to the distance calculation.

C.7 Semantic and Gallery-Size Analyses

Table 19 reports each text encoder’s highest correlation with the human SICK relatedness scores and its calibrated speech-text alignment. Semantic performance ranges from 0.744 to 0.891, but it is not associated with calibrated mutual k NN across models ($\rho = 0.103$, $p = 0.777$). The corresponding association is also absent for raw mutual k NN ($\rho = 0.055$, $p = 0.881$) and calibrated RBF-CKA ($\rho = -0.430$, $p = 0.214$).

Table 20 reports sample-size sensitivity. At fixed $k = 10$, raw overlap decreases as the comparison set grows, while chance adjustment removes most of this change. When k is proportional to the number of samples, the raw score does not exhibit the same decline.

Finally, Table 21 compares cross-modal alignment with the stability of a speech encoder’s own

Metric	Median layer		Text-layer frequency	
	Speech	Text	Layer 0	Layers 0–2
Mutual k NN, $k = 10$	20	2	31.1%	51.3%
CKNNA, $k = 10$	20	0	72.2%	85.6%
Linear CKA	22	4	37.4%	46.3%
RSA, Spearman	20	4	7.3%	28.1%
RBF-CKA, $\sigma = 0.1$	24	0	62.7%	71.8%
RBF-CKA, $\sigma = 0.5$	23	0	51.3%	52.1%
RBF-CKA, $\sigma = 2$	20	4	31.6%	40.3%
RBF-CKA, $\sigma = 5$	20	4	29.4%	39.0%

Table 10: Locations of the maximum raw layer-pair scores. Text-layer frequencies give the percentage of conditions in which the maximum uses the input embedding output or one of the first three extracted text representations.

Component	df	Sum of squares	
		($\times 10^{-3}$)	Share
Speech encoder	8	102.724	95.356%
Text encoder	9	3.502	3.251%
Speech $\times$ text	72	1.501	1.393%
Total	89	107.727	100.000%

Table 11: Descriptive decomposition of variation among the 90 raw mutual k NN encoder-pair means. Each pair-level mean summarizes 24 speaker-emotion conditions.

neighborhoods when the same sentences are realized by different speakers or emotions. The comparison is descriptive because the cross-modal and within-speech scores are produced by different pairing procedures.

C.8 Model rankings are stable after calibration but depend on the alignment metric

Aggregation-aware calibration has little effect on the relative ordering of models under mutual k NN. HuBERT X-Large produces the highest calibrated alignment among speech encoders at 0.283, followed by Whisper Medium at 0.271 and HuBERT Large at 0.264. wav2vec 2.0 Base produces the lowest value at 0.173. The raw and calibrated speech-encoder rankings are identical ($\rho = 1.000$), because calibration reduces each encoder’s mean score by a similar amount, between 0.017 and 0.019. Text-encoder rankings are also nearly unchanged ($\rho = 0.988$): E5-v2 Base remains first at 0.242, whereas Qwen3-Embedding 8B remains last at 0.219, with only the two GTE models exchanging adjacent positions. Figure 3 shows both rankings before and after calibration, and complete values are reported in Appendix C.4.

Rankings differ more substantially across alignment metrics. On the speech side, mutual k NN

and linear CKA produce broadly similar orderings ($\rho = 0.700$, $p = 0.036$), although the Whisper encoders generally rank higher under linear CKA than under local neighborhood alignment. On the text side, however, the mutual k NN ranking is not associated with either the linear CKA ranking ($\rho = -0.212$, $p = 0.556$) or the RSA ranking ($\rho = 0.127$, $p = 0.726$). For example, Qwen3-Embedding 0.6B ranks sixth under mutual k NN but first under linear CKA. At the level of the 90 speech-text encoder pairs, the five highest-scoring pairs under mutual k NN have no overlap with the five highest-scoring pairs under linear CKA. Claims about which models align most strongly must therefore specify whether alignment refers to local neighborhood structure or global geometry.

Speech encoder	Raw	Calibrated	Reduction	Raw rank	Cal. rank
HuBERT X-Large	0.300 ± 0.010	0.283 ± 0.011	0.017	1	1
Whisper Medium	0.288 ± 0.012	0.271 ± 0.012	0.017	2	2
HuBERT Large	0.281 ± 0.008	0.264 ± 0.009	0.017	3	3
WavLM Large	0.259 ± 0.007	0.241 ± 0.007	0.018	4	4
Whisper Large-v3	0.247 ± 0.009	0.229 ± 0.010	0.018	5	5
WavLM Base+	0.232 ± 0.005	0.214 ± 0.005	0.018	6	6
Whisper Small	0.222 ± 0.007	0.204 ± 0.008	0.018	7	7
wav2vec 2.0 Large	0.221 ± 0.004	0.203 ± 0.004	0.019	8	8
wav2vec 2.0 Base	0.192 ± 0.004	0.173 ± 0.004	0.019	9	9

Table 12: Mutual k NN alignment by speech encoder. Values are mean $\pm$ standard deviation across the 10 text-encoder means. Reduction is the raw mean minus the calibrated mean. Calibration reduces every speech-encoder mean by 0.017–0.019 without changing the ordering.

Text encoder	Raw	Calibrated	Reduction	Raw rank	Cal. rank	Δ rank
E5-v2 Base	0.260 ± 0.039	0.242 ± 0.040	0.018	1	1	0
E5-v2 Large	0.256 ± 0.038	0.238 ± 0.039	0.018	2	2	0
BERT Large	0.253 ± 0.036	0.235 ± 0.037	0.018	3	3	0
BERT Base	0.252 ± 0.038	0.235 ± 0.039	0.018	4	4	0
E5-v2 Small	0.252 ± 0.039	0.234 ± 0.040	0.018	5	5	0
Qwen3-Embedding 0.6B	0.247 ± 0.033	0.229 ± 0.034	0.018	6	6	0
GTE-v1.5 Base	0.245 ± 0.036	0.227 ± 0.037	0.018	8	7	−1
GTE-v1.5 Large	0.245 ± 0.036	0.227 ± 0.037	0.018	7	8	+1
Qwen3-Embedding 4B	0.243 ± 0.030	0.225 ± 0.031	0.018	9	9	0
Qwen3-Embedding 8B	0.238 ± 0.033	0.219 ± 0.034	0.018	10	10	0

Table 13: Mutual k NN alignment by text encoder. Values are mean $\pm$ standard deviation across the 9 speech-encoder means. Calibration leaves the ordering almost unchanged; only the two adjacent GTE models exchange positions.

Modality	Benchmark	n	ρ	p
Text	MTEB	10	−0.860	0.007
Speech	SUPERB	5	+0.700	0.19

Table 14: Association between calibrated mutual k NN alignment and external capability evaluations. The speech analysis includes only the five self-supervised encoders with comparable SUPERB evaluations.

Similarity measure	Spearman		Pearson	
	ρ	p	r	p
Mutual k NN, $k = 10$	0.139	0.621	0.001	0.996
Linear CKA	0.007	0.980	−0.216	0.440
RSA, Spearman	0.221	0.428	−0.120	0.669

Table 15: Association between ECAPA2 speaker distance and representation similarity across the 15 unordered speaker pairs. The pairs reuse the same six speakers and are therefore not independent; the statistics are descriptive.

Dependent variable	All conditions		Non-neutral only	
	ρ	p	ρ	p
Calibrated mutual k NN	−0.189	0.376	−0.044	0.861
Selected speech depth	0.077	0.721	0.127	0.616
Selected text depth	−0.154	0.473	−0.740	< 0.001

Table 16: Spearman associations with prosodic distance. The complete analysis contains 24 speaker-emotion conditions; the non-neutral analysis contains 18. Depth is expressed as a percentage of the available representations in the corresponding model.

Emotion	d_{pros}	Calibrated mutual $k\text{NN}$	Speech depth (%)	Text depth (%)
Angry	6.812 ± 0.493	0.243 ± 0.051	80.012	12.597
Happy	5.339 ± 1.458	0.208 ± 0.013	79.796	15.176
Neutral	0.000	0.246 ± 0.047	79.888	13.213
Sad	3.785 ± 0.922	0.227 ± 0.029	79.836	16.231

Table 17: Prosodic distance, calibrated alignment, and selected relative layer depths by intended emotion. Values are mean $\pm$ standard deviation across the six speakers where applicable. Neutral distance is zero by construction.

Feature	ρ	p
Mean intensity	0.638	< 0.001
Mean F0	0.608	< 0.001
Median F0	0.562	< 0.001
F0 standard deviation	0.472	< 0.001
F0-confidence standard deviation	-0.406	< 0.001
Mean voiced-frame F0 confidence	0.392	< 0.001
Rate of F0 change	0.283	0.017

Table 18: Spearman associations between the absolute standardized feature values and prosodic distance across 71 non-neutral speaker-emotion-system observations. Only features with $p < 0.05$ are shown.

Text encoder	Best layer	SICK relatedness ρ	Calibrated mutual k NN
GTE-v1.5 Base	12	0.891	0.227
E5-v2 Small	12	0.889	0.234
E5-v2 Base	12	0.888	0.242
E5-v2 Large	24	0.887	0.238
GTE-v1.5 Large	24	0.881	0.227
Qwen3-Embedding 4B	36	0.817	0.225
Qwen3-Embedding 0.6B	28	0.795	0.229
Qwen3-Embedding 8B	36	0.784	0.219
BERT Large	6	0.746	0.235
BERT Base	1	0.744	0.235

Table 19: Sentence-level semantic performance and speech-text alignment by text encoder. Best layer is selected independently by maximizing the Spearman correlation between embedding cosine similarity and the human relatedness judgments for the 300 selected SICK pairs.

N	Fixed neighborhood			Proportional neighborhood		
	k	Raw	Adjusted	k	Raw	Adjusted
128	10	0.299	0.239	1	0.245	0.239
256	10	0.264	0.234	3	0.249	0.240
512	10	0.249	0.234	5	0.252	0.245
600	10	0.249	0.236	6	0.253	0.246

Table 20: Mutual k NN sensitivity to comparison-set size. Values are maximum layer-pair scores averaged across the 90 encoder pairs. Results for $N < 600$ are averaged over 10 random subsets. Adjusted scores account for the changing random-neighborhood baseline.

Comparison	Mean mutual k NN
Speech across emotions	0.152
Speech across speakers	0.156
Speech-text	0.249

Table 21: Neighborhood agreement when sentence content is held constant. Within-speech rows compare representations from the same speech encoder across acoustic realizations; the speech-text row reports the mean raw cross-modal alignment.