Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations

Mukhtar Mohamed, Oli Danyi Liu, Hao Tang, Sharon Goldwater

University of Edinburgh, United Kingdom

mukhtaralgezoli@gmail.com, oli.liu@ed.ac.uk, hao.tang@ed.ac.uk, sgwater@inf.ed.ac.uk

Abstract

Self-supervised speech representations can hugely benefit downstream speech technologies, yet the properties that make them useful are still poorly understood. Two candidate properties related to the geometry of the representation space have been hypothesized to correlate well with downstream tasks: (1) the degree of orthogonality between the subspaces spanned by the speaker centroids and phone centroids, and (2) the isotropy of the space, i.e., the degree to which all dimensions are effectively utilized. To study them, we introduce a new measure, Cumulative Residual Variance (CRV), which can be used to assess both properties. Using linear classifiers for speaker and phone ID to probe the representations of six different self-supervised models and two untrained baselines, we ask whether either orthogonality or isotropy correlate with linear probing accuracy. We find that both measures correlate with phonetic probing accuracy, though our results on isotropy are more nuanced.

Index Terms: model analysis, representational geometry

1. Introduction

Self-supervised speech representations have made a huge impact on downstream speech technologies, yet the properties that make their representations useful are still poorly understood. Benchmarks indicate that both phone and speaker labels are, to a large degree, linearly separable in the representations of popular recent models [1], and beyond this, a number of studies have compared the extent to which these labels are recoverable from the representations of different models [1, 2] or across different layers of the same model [3, 4]. However, these analyses say little about *how* such information is represented, beyond just assessing the linear separability of classes. Here, we address this question using a *geometric* approach—an approach that is widely used for analyzing self-supervised models of text (e.g., [5–10]) as well as high-dimensional brain imaging data (e.g., [11–13]), but has received only a little attention in the speech technology community [14–17].

To assist our analysis, we develop a new measure for analyzing high-dimensional distributions, the Cumulative Residual Variance (CRV). When applied to datasets X and Y embedded in the same high dimensional space, the CRV of X with respect to Y, denoted $X\backslash Y$, provides a quantitative measure of the degree to which the principal components of Y are orthogonal to those of X. Meanwhile, $X\backslash X$ is a measure of the *isotropy* of X—the degree to which X effectively utilizes all dimensions of the embedding space, i.e., has uniform covariance [18].

Using this measure, we draw on two previous lines of work that suggest potentially fruitful analyses. First, we build on a recent study which analyzed LSTM models trained using two different loss functions and demonstrated that speaker and phonetic information were represented in orthogonal subspaces [17]. The CRV measure allows us to better quantify orthogonality, and we use it to analyze several additional models with a variety of architectures, loss functions, and training data. In experiments on English LibriSpeech, we show that, unlike randomly initialized (untrained) models, all trained models have a high degree of orthogonality between the speaker and phonetic subspaces. In addition, for all six trained models, the accuracy of a phone classifier trained on the model representations is significantly correlated with the CRV between the two subspaces.

Next, we explore whether and how the isotropy of the representational space might predict phone or speaker classification accuracy. It has been argued in the NLP literature that higher isotropy is desirable in an embedding space (e.g., [6, 9] and see review in [19]). However, we did not find strong evidence for this hypothesis: when we computed the isotropy and phone (or speaker) classification accuracy for different layers of each model, we found a statistically significant correlation in only two out of six trained models. On the other hand, we did find a strong and consistent correlation between phone classification accuracy and the isotropy of the phone class centroids. This suggests that having evenly distributed centroids is more important for classification accuracy in these models than having evenly distributed frame representations.

2. Isotropy and orthogonality

In NLP, most researchers have argued that representations with greater isotropy are desirable [6,9,10,20]; but see [21]. However, Rudman *et al.* [18] noted that the measures of "isotropy" used in much of that work do not match its mathematical definition—that is, the extent to which the covariance matrix is proportional to the identity matrix. They introduced (and demonstrated the correctness of) a new measure called IsoScore, and later used it to show that isotropy is in fact *negatively* correlated with task performance in several BERT models [19]. Meanwhile, we know of only one study of isotropy in models of speech [22], which found a strong positive correlation between IsoScore and word discrimination performance in supervised acoustic word embedding models. Here, we explore whether isotropy can predict either phone or speaker classification performance in self-supervised representations.

As noted above, IsoScore [18] is one way to measure isotropy. IsoScore ranges from 0 (minimally isotropic) to 1 (maximally isotropic), and can be interpreted as the approximate proportion of the dimensions that are uniformly utilized. Computing the IsoScore for a point cloud $X \subseteq \mathbb{R}^d$ starts by applying PCA, then finding the Euclidean distance between the length-normalized vector of eigenvalues Λ (the diagonal of the covariance matrix) and the diagonal of the identity matrix in $\mathbb{R}^d$. This distance is then normalized and rescaled to fall between 0 and 1.

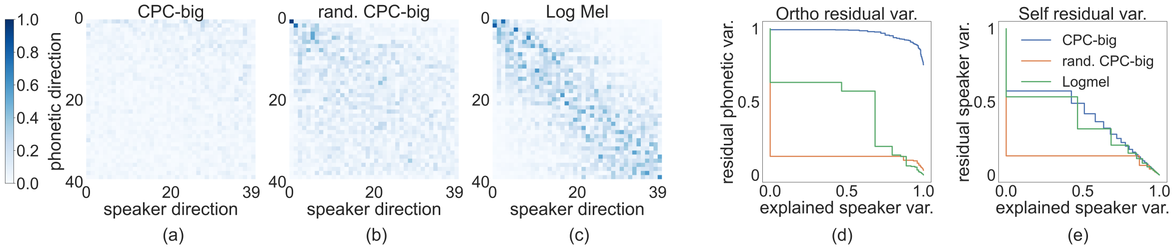

Figure 1: *Evaluating orthogonality as cosine similarities between principal components (a-c) versus using residual variance (d). Isotropy can be evaluated with self residual variance (e).*

While IsoScore has properties that can be desirable (e.g., it allows direct comparisons between spaces of different dimensionality on the same 0-1 scale), it is not the only possible measure of isotropy. For example, Del Giudice [23] discusses estimators of "Effective Dimensionality" which normalize Λ to create a probability distribution, then calculate its entropy to measure deviance from uniformity, and return a value interpreted as the *number* (rather than *proportion*) of dimensions uniformly utilized.

Apart from isotropy, orthogonality is also a desirable property when learning representations, since encoding different kinds of information in orthogonal dimensions or subspaces would allow them to be easily disentangled. In fact, there have been attempts in representation learning to enforce such orthogonality to enable disentanglement [24, 25]. There is also evidence that human brains encode different aspects of the same item in orthogonal coding axes, thereby minimizing interference and maximizing robustness [26, 27]. However, [17] (henceforth, Liu *et al.*) is the only work we know of to explore orthogonality in either supervised or unsupervised speech models. We describe their method, and how we build on it, in more detail below.

3. Measuring orthogonality

Before evaluating the orthogonality between speaker and phonetic encoding, we first follow Liu *et al.* in identifying speaker directions and phonetic directions. Phonetic directions are found by aggregating the frame-level representations for each of the 39 phones (based on forced alignment) to obtain their centroids, and then applying principal component analysis to the centroids. The 39 principal components found represent the *phonetic directions*, along which the variance between the centroids is maximized. The same method is used to obtain speaker directions using the speakers in the dataset.

Our next step diverges from Liu *et al.*'s: while they looked at the cosine similarities between the phonetic and speaker directions (§3.1), we propose a new measure that quantifies orthogonality with a single numerical value (§3.2).

3.1. Cosine similarity between principal directions

For each pair of a speaker and a phonetic direction, Liu *et al.* computed the degree of orthogonality by taking the absolute value of their cosine similarity, with 0 being perfectly orthogonal and 1 being perfectly aligned. Figs. 1a-c present the pairwise similarity for representations extracted from (a) the second LSTM layer of the same CPC-big model used by Liu *et al.* (from [28]); (b) the same layer of a randomly initialized CPC model that has not been trained, and (c) log Mel features. Confirming Liu *et al.*'s results, Fig. 1a shows very low similarities between any pair of phonetic and speaker directions, indicating the two types of information are largely encoded orthogonally.

While the similarity matrix gives some indication of the relationship between the directions encoding speaker and phonetic information, it can be difficult to summarize with a single number: we need to consider the degree of alignment between every pair of directions in order to fully capture the degree of orthogonality between speaker and phonetic encoding. In addition, alignment between principal directions with large eigenvalues means overall lower orthogonality than alignment between principal directions with small eigenvalues, but the matrix does not reflect the amounts of variance in each principal direction.

3.2. Cumulative Residual Variance (CRV)

We propose Cumulative Residual Variance (CRV) as a quantitative measure of orthogonality between datasets X and Y embedded in $\mathbb{R}^d$. CRV satisfies the two desiderata mentioned above: (1) it captures the interaction between every pair of principal directions and (2) it weights the contribution from each principal direction in proportion to its relative explained variance. Here, we set X and Y to be the speaker and phone centroids (or vice versa), so the number of data points n_X and n_Y is less than the dimensionality d, and each dataset only spans a subspace of $\mathbb{R}^d$. However, this need not be true in general; for example CRV could be applied to the sets of frame-level representations from two different speakers or two different phones.

In short, the CRV of Y with respect to X, written as $Y\backslash X$, evaluates how much variance is preserved in Y as the principal directions of X are collapsed one by one.[1] As in Liu *et al.*, "collapsing" a direction v from a dataset Y refers to the operation of projecting Y onto the subspace orthogonal to v, i.e. $Y' = Y - (Yv)v^\top$. Collapsing v affects any principal direction of Y that is not orthogonal to v, which addresses the first desideratum. We evaluate the effect of the collapsing operation by computing the residual variance in Y', as given by PCA. The larger the residual variance Y' has, the more orthogonal Y is to v.

The residual variances computed in this way can be plotted as in Fig. 1d, where for any given x-axis value, its y value is the proportion of variance remaining in Y after collapsing the minimum number of top principal directions of X such that at least x proportion of X's variance has been removed. CRV is then computed from this plot as the area under the curve (AUC), to yield a single numerical value. In this way, the effect of collapsing each direction is weighted by the variance explained by that direction, hence CRV also satisfies our second desideratum.

In Fig. 1d, we plot residual variance of the phone centroids with respect to the explained variance in the speaker centroids for representations from a trained and an untrained CPC[2] as well as for log Mel features. We can see that the relative magnitude of the AUC is CPC, followed by log Mel and untrained CPC. While the strong orthogonality in the trained CPC is consistent with

[1] Note that CRV is an asymmetrical distance measure. Like KL divergence, it could be symmetrised as $X\backslash Y + Y\backslash X$, if desired.

[2] Though CPC is a loss function, with a slight abuse, we refer to a randomly initialized CPC-big in [28] as untrained CPC.

Fig. 1a, the relative degree of orthogonality between log Mel and untrained CPC is less salient from Fig. 1b-c. There are more dark spots in Fig. 1c, indicating more pairs of aligned speaker and phonetic directions in log Mel, but this should have less effect on overall orthogonality as compared to the top left corner of Fig. 1b, which shows that the first two speaker and phonetic directions of the untrained CPC are very strongly aligned. This is properly reflected in Fig. 1d.

3.3. Evaluating isotropy with Self-CRV

A byproduct of CRV is *self-CRV*, or $Y \backslash Y$, which evaluates the degree of isotropy of Y in $\mathbb{R}^d$. If Y is highly anisotropic, its variance will be concentrated around a few directions. This results in a residual variance curve with a small AUC, as illustrated in Fig. 1e for untrained CPC.

Self-CRV is closely related to IsoScore, both being functions of the eigenvalues of the dataset. However, IsoScore measures isotropy as a percentage of representation dimensions, whereas self-CRV accounts for the absolute number of isotropic dimensions and is comparable across models with different dimensions as long as the number of data points in Y is smaller than the dimensionality of all models (as in our subspace analyses). After multiplying IsoScore by model dimension, we found a Spearman's rank correlation of 1 between it and self-CRV.

4. Experimental Setup

Models In addition to CPC-big, we measured orthogonality and isotropy in five pre-trained Transformer-based English self-supervised speech models: HuBERT (base-ls960) [29], wav2vec 2.0 (base-960h) [30], WavLM (base) [31], WavLM+ (base-plus) [31], and Data2Vec (base-960h) [32]. Apart from the architecture, CPC-big differs from the Transformer-based models in its dimensionality (512 vs. 768), number of layers (5 CNN followed by 4 LSTM vs. 7 CNN followed by 12 Transformer blocks), frame rate (10ms vs. 20ms) and amount of training data (6k hr vs. 960 hr for all others except WavLM+, which used 96k hr). To determine the degree of orthogonality and isotropy in these models before training, we also tested representations extracted from a HuBERT model and a CPC-big model with just random initialization and no training. Since the Transformer models we tested have mostly the same architecture and are distinguished by the training methods and objective, the untrained HuBERT is representative of the other Transformer models. Finally, 40-dimensional log Mel features are used as a baseline.

Dataset We perform our analysis on the dev-clean subset of LibriSpeech [33], which matches the language (English) and genre (read speech) of the pre-trained models and was also used in Liu *et al.*[3] Dev-clean contains 40 speakers, each contributing at least eight minutes of speech. We used half of dev-clean for training classifiers and half for testing, with different splits depending on the scenario, as described below.

Probing classifiers Our analysis focuses on speaker information and phonetic information, due to their influence on a variety of downstream speech tasks. We train logistic regression classifiers to predict the speaker (or phone) label based on a single representation frame. In previous work, frames are typically pooled across phones [4, 34] or utterances (for speaker ID) [1]; but like Liu *et al.* we use individual frames, so we can analyze how both types of information sit in the same set of embeddings.

For speaker classification, we obtain speaker labels from the LibriSpeech metadata and train the probing classifier on a random half of each speaker's utterances, using the other half for testing. For phone classification, we obtain the phone labels from forced alignments with Kaldi. We evaluated phone accuracy in two ways: *shared speakers* (as in Liu *et al.*), where the same speakers appear in both training and test, and the more standard *across-speaker*, where we trained on data from a random half of the speakers and tested on the other half. In practice, the measures are very strongly correlated and don't differ much, so in this paper we only report across-speaker phone accuracy.

Computing CRV, IsoScore, and correlations We computed CRV and IsoScore for each layer of each model by first encoding the utterances from LibriSpeech dev-clean to obtain the representations. We then computed the phone and speaker centroids and CRV values as described in §3. In particular, Ph\Spk measures the orthogonality of the phone and speaker subspaces, and Ph\Ph and Spk\Spk measure the isotropy of the phone and speaker subspaces, respectively. We computed the IsoScore using a random sample of 250,000 frames. Finally, for correlations between classifier accuracy and CRV or IsoScore, we computed Spearman (rank) correlation, since it is less sensitive to outliers and we have no reason to believe that correlations will be linear.

5. Results and Discussion

5.1. Layerwise Classification Accuracy

Fig. 2 (lst column) shows the results of our probing classifiers for phones (top) and speakers (bottom), across all layers of each model[4]. For phones, our findings align closely with those of [3, 4, 34], despite analyzing frame-wise rather than pooling the representations for each phone token. That is, for wav2vec2 (and data2vec) the highest probing accuracies are in the late middle layers, while for HuBERT-family models (HuBERT, WavLM, WavLM+), accuracy remains high through the final layers.

To the best of our knowledge, previous studies have only reported speaker probing accuracy across all the layers of HuBERT [35]. Extending the layerwise analysis of speaker information to the other widely-used SSL models, we find far more variation here than with phone accuracy, perhaps because all of these models, despite being self-supervised, are designed with ASR in mind. We see especially poor linear separability in the later layers of wav2vec2 and data2vec, where speaker accuracy is even worse than the randomly initialized ones. We speculate that the rising pattern of speaker accuracy in the randomly initialized models may be because the model incorporates more context in later layers, allowing the model to effectively average features over the whole utterance.

5.2. Geometry of the phone and speaker subspaces

Layer-wise CRV and self-CRV results for all models are shown in Fig. 2, columns 2 and 3. Like the CPC model studied by Liu *et al.*, all trained Transformer models have high Ph\Spk orthogonality. Interestingly, untrained HuBERT (unlike untrained CPC) also reaches a somewhat high Ph\Spk value in the final layers, although still lower than the trained models. The trained models also show high isotropy in the phone and speaker centroids (Ph\Ph and Spk\Spk), though as with probing accuracy, the difference between trained and untrained models is much more

[3] We hope in future to examine how much these results generalize, by extending the analyses to other genres and languages, either using different pre-trained models or by testing these models on other data.

[4] Due to space constraints, not all results can be displayed in the paper. The complete spreadsheet of our results, and the code for computing CRV can be found at `https://github.com/uililo/cumulative-residual-variance`.

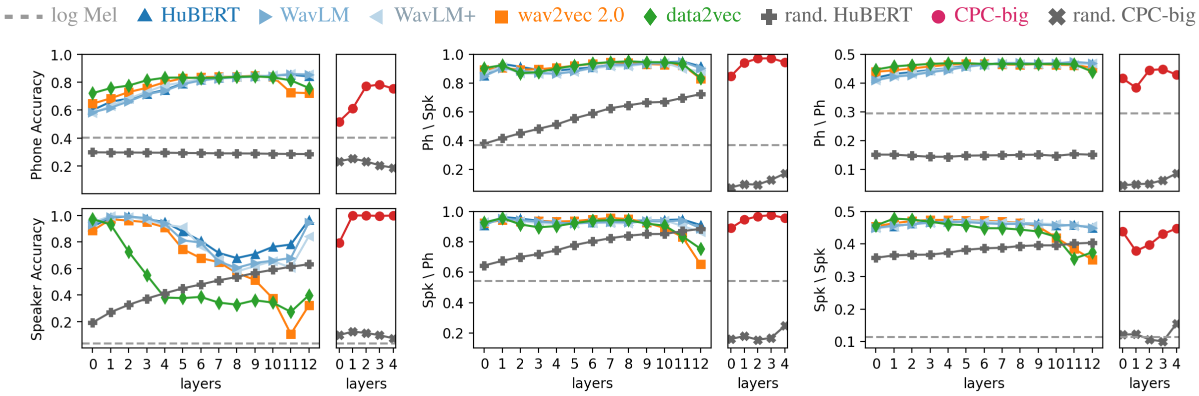

Figure 2: *Layerwise results for all models, showing (in columns from left to right): Phone and speaker classiciation accuracy; CRV orthogonality measures (Ph\Spk and Spk\Ph); and self-CRV measures (Ph\Ph and Spk\Spk).*

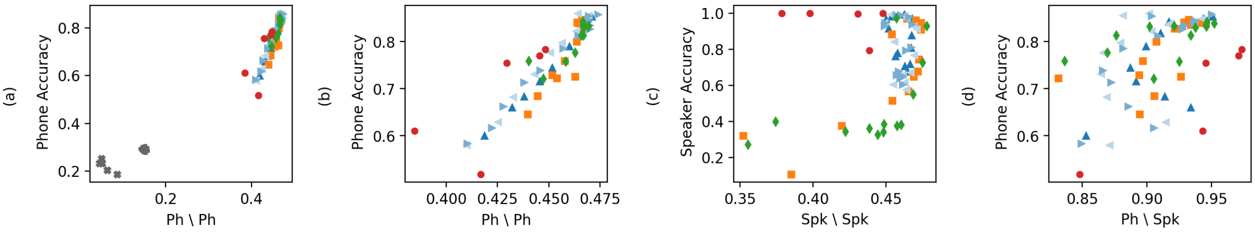

Figure 3: *Correlations between orthogonality or isotropy measures and probing accuracies. Marker styles are as in Fig. 2.*

striking for the phonetic measure, suggesting that model training reorganizes the representational geometry of the phonetic information more than the speaker information.

We then computed rank correlations ρ between each of the four CRV measures and the speaker or phone accuracies. The most striking correlation is between Ph\Ph and phone accuracy, as shown in Figs. 3a (all models) and 3b (trained models only). Pooling all datapoints together, $\rho = 0.94$, and the trained models individually each have ρ from 0.69 to 0.9 (all values $p < 0.05$). In contrast, we only found statistically significant correlations between Spk\Spk and speaker accuracy (Fig. 3c) in wav2vec2 and data2vec, and no significant correlation when pooling the results from all trained models.

For the orthogonality measures, we found significant correlations between Ph\Spk and phone accuracy (Fig. 3d) in each of the trained models ($\rho = 0.54$-0.78 for Transformer models, 1.0 for CPC), as well as in the pooled data ($\rho = 0.54$), although the correlations are weaker than for Ph\Ph. Correlations between Spk\Ph and speaker accuracy are even weaker, reaching significance on the pooled data, but not for any individual model.

Altogether, our results suport Liu *et al.*'s claim that orthogonality between the phonetic and speaker subspace is relevant for extracting phonetic information, but also suggest that isotropy of the phonetic space may be even more critical. It is less clear why the geometry of speaker information is less correlated to speaker classification, and to what extent this result is due to model training that is implicitly focused on ASR performance.

5.3. Isotropy of the frame representation space

Finally, we evaluated the isotropy of frame representations (rather than the centroids). For this, we used IsoScore, which (1) has a rank correlation of 1 with self-CRV as isotropy measures of speaker or phone centroids, and (2) is easier to compute than self-CRV when applied to a large number of representations. The IsoScore values were low, ranging from 0.18 to near 0 across models and layers, similar to the range found by Rudman *et al.* [18] for contextualized word embedding models. Also, the IsoScore values for untrained HuBERT were comparable to those of the trained models. We find a statistically significant ($p < 0.05$) positive correlation with phone probing accuracy in HuBERT and WavLM, and when pooling results from all trained models; but for speaker probing accuracy we found negative correlations in the same two models, and no significant pooled correlation. These mixed results suggest that isotropy of the representation space itself is not necessarily a good predictor of model performance, especially if different tasks are considered.

6. Conclusion

This paper introduced the Cumulative Residual Variance as a new way to analyze the representational geometry of high-dimensional spaces, and used CRV and IsoScore to examine whether orthogonality or isotropy can predict phone or speaker probing accuracy in self-supervised speech models. We did not find strong evidence that isotropy of the frame representations is meaningful, but we did show that phone probing accuracy is correlated with the degree of orthogonality between the subspaces defined by the phone and speaker centroids, and even more strongly with the isotropy of the phone centroids themselves. These findings suggest that geometric analyses may be a productive route for future study, particularly if they can be more closely connected to theoretical analyses such as those of [13]. For instance, [13] highlights the relevance of four different geometric properties, including the distance between class centroids (related to our subspace isotropy measure) as well as the isotropy of the individual class manifolds (i.e., phones or speakers). We hope that our work may inspire further exploration of these connections.

7. Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, and also received support from Google.

8. References

[1] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021.

[2] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[3] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[4] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[5] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics*, 2013, pp. 746–751.

[6] X. Cai, J. Huang, Y. Bian, and K. Church, "Isotropy in the contextual embedding space: Clusters and manifolds," in *International conference on learning representations*, 2020.

[7] T. A. Chang, Z. Tu, and B. K. Bergen, "The geometry of multilingual language model representations," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022.

[8] E. Hernandez and J. Andreas, "The low-dimensional linear geometry of contextualized word representations," *Proceedings of the 25th Conference on Computational Natural Language Learning*, 2021.

[9] J. Mu, S. Bhat, and P. Viswanath, "All-but-the-top: Simple and effective postprocessing for word representations," *Proceedings of the International Conference on Learning Representations*, 2017.

[10] W. Timkey and M. Van Schijndel, "All bark and no bite: Rogue dimensions in transformer language models obscure representational quality," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[11] N. Kriegeskorte and R. A. Kievit, "Representational geometry: integrating cognition, computation, and the brain," *Trends in cognitive sciences*, vol. 17, no. 8, pp. 401–412, 2013.

[12] S. Saxena and J. P. Cunningham, "Towards the neural population doctrine," *Current opinion in neurobiology*, vol. 55, pp. 103–111, 2019.

[13] B. Sorscher, S. Ganguli, and H. Sompolinsky, "Neural representational geometry underlies few-shot concept learning," *Proceedings of the National Academy of Sciences*, 2022.

[14] C. Stephenson, J. Feather, S. Padhy, O. Elibol, H. Tang, J. McDermott, and S. Chung, "Untangling in invariant speech recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[15] B. M. Abdullah, M. Mosbach, I. Zaitova, B. Möbius, and D. Klakow, "Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study," in *Proc. Interspeech*, 2021.

[16] K. Choi and E. J. Yeo, "Opening the black box of wav2vec feature encoder," *arXiv preprint*, 2022.

[17] O. D. Liu, H. Tang, and S. Goldwater, "Self-supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces," in *Proc. Interspeech*, 2023.

[18] W. Rudman, N. Gillman, T. Rayne, and C. Eickhoff, "Isoscore: Measuring the uniformity of embedding space utilization," *Findings of the Association for Computational Linguistics*, 2022.

[19] W. Rudman and C. Eickhoff, "Stable anisotropic regularization," *International Conference on Learning Representations*, 2024.

[20] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[21] Y. Ding, K. Martinkus, D. Pascual, S. Clematide, and R. Wattenhofer, "On isotropy calibration of transformers," *Proceedings of Workshop on Insights from Negative Results in NLP*, 2021.

[22] B. M. Abdullah and D. Klakow, "Analyzing the representational geometry of acoustic word embeddings," *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2022.

[23] M. Del Giudice, "Effective dimensionality: A tutorial," *Multivariate behavioral research*, vol. 56, no. 3, pp. 527–542, 2021.

[24] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "Fairness by learning orthogonal disentangled representations," in *European Conference on Computer Vision*. Springer, 2020.

[25] J. Cha and J. Thiyagalingam, "Orthogonality-enforced latent space in autoencoders: An approach to learning disentangled representations," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[26] T. Flesch, K. Juechems, T. Dumbalska, A. Saxe, and C. Summerfield, "Orthogonal representations for robust context-dependent task performance in brains and neural networks," *Neuron*, vol. 110, no. 7, 2022.

[27] A. Libby and T. J. Buschman, "Rotational dynamics reduce interference between sensory and memory representations," *Nature neuroscience*, vol. 24, no. 5, pp. 715–726, 2021.

[28] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," *arXiv preprint arXiv:2011.11588*, 2020.

[29] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, 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.

[30] 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.

[31] 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*, 2022.

[32] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015.

[34] P. C. English, J. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2022.

[35] H.-J. Chang, A. H. Liu, and J. Glass, "Self-supervised Fine-tuning for Improved Content Representations by Speaker-invariant Clustering," in *Proc. INTERSPEECH 2023*, 2023, pp. 2983–2987.