Probing emergent geometry in speech models via replica theory

Anonymous authors

ABSTRACT

The success of deep neural networks in visual tasks have motivated recent theoretical and empirical work to understand how these networks operate. Meanwhile, deep neural networks have also achieved impressive performance in audio processing applications, both as sub-components of larger systems or as complete end-to-end systems. In this work, we employ a recently developed statistical mechanical theory that connects geometric properties of network representations with class separability to probe how information is untangled within neural networks trained to recognize speech. We find that speech recognition models carry out significant layerwise and temporal untangling of words by efficiently extracting task-relevant features. This untangling results from a decrease in the per-class radius and dimension, and a reduction in the correlation between class centers.

1 Introduction

Understanding invariant object recognition is one of the key challenges in artificial intelligence. When objects exhibits stimulus variability, the set of different representations corresponding to the same object category form an object manifold. In vision systems, it has been hypothesized that these "object manifolds", hopelessly entangled in the input, become "untangled" across the visual hierarchy both in the brain and in deep neural networks[1;2]. Auditory recognition also requires the separation of highly variable inputs according to category, and could also involve the untangling of 'auditory class manifolds'.

In recent years, hierarchical neural network models have achieved state of the art performance in automatic speech recognition (ASR)[3;4]. Understanding how these end-to-end models represent speech information remains a major challenge[5;6]. Several studies have analyzed how phonetic information is encoded in acoustic models[7;8;9], and how it is embedded across layers by making use of classifiers[10;11;12;5]. However, whether such representations also encode higher-level concepts such as words is unknown. Deep neural networks trained on speech recognition also resemble human behavior and auditory cortex activity[13]. Ultimately, understanding speech-processing in deep networks may also shed light on understanding how the brain processes auditory information.

Prior work characterizes how object representations become more linearly separable across visual hierarchy in biological systems[1]. Representations have also been compared across different networks, layers, and training epochs using Canonical Correlation Analysis (CCA)[14;15;16], and Representational similarity analysis (RSA)[17;18]. Explicit geometric measures have been used to understand deep networks, such as curvature[2;19], geodesics[20], and Gaussian mean width[21]. However, none of these measures make a theoretical connection between the separability of object representations and their geometrical properties.

In this work, we employ a recently developed mean-field theoretical framework[22;23;24] based on replica method[25;26;27] that links the geometry of object manifolds to the capacity of a linear classifier in order to quantify the information stored about object categories per feature dimension. This method has been previously used to undersatnd object untangling in visual CNNs[24]. Here we apply manifold analyses to auditory models for the first time, and show that neural network speech recognition systems also 'untangle' words, even when trained only for character-level output.

2 Methods

We apply the Mean-Field Theory (hereafter, MFT) based manifold analysis technique[23;24] on features extracted from each network layer. Formally, if we have P object manifolds (e.g. words), we can construct a dataset with pairs (x_i, y_i), where x_i is the auditory input, and $y_i \in P$ denotes the object manifold. For each manifold p, we extract $Net_t^l(x)$, the output of the network at time t in layer l, for all input x whose corresponding label is the p^{th} manifold. We analyze these activations to compute the manifold capacity, dimension, radius, and correlations between manifolds.

The manifold capacity obtained by this analysis technique captures the linear separability of object manifolds (See Appendix Fig.A1-A2). Furthermore, recent work[23;24] shows that manifold capacity is tightly connected to the size and dimensionality of the manifolds, implying that the representation separability can be understood geometrically using MFT techniques. We measure these properties for word categories, which allows us to quantify the amount of invariant information about each word and the characteristics of the emergent representation learned by the speech models.

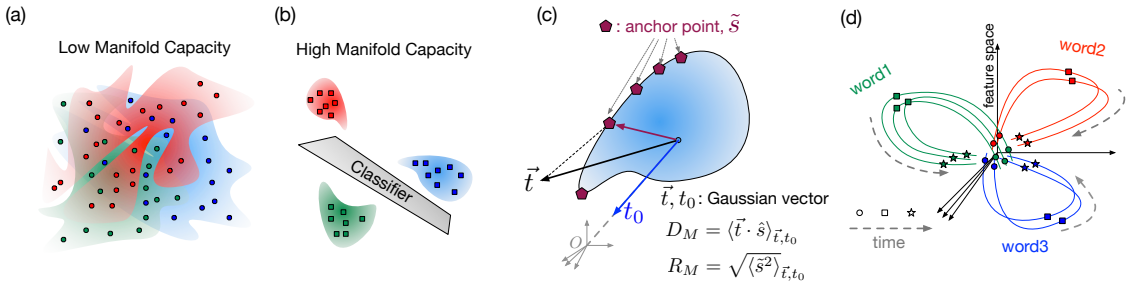

Figure 1: **Illustration of word manifolds.** (a) highly tangled manifolds, in low capacity regime (b) untangled manifolds, in high capacity regime (c) Manifold Dimension captures the projection of a Gaussian vector onto the direction of an anchor point, and Manifold Radius captures the norm of an anchor point in manifold subspace. (d) Illustration of untanglement of words over time

2.1 Object Manifold Capacity and Mean Field Theoretic manifold analysis

In a system where P object manifolds are represented in N ambient dimensions, the 'load' in the system is defined by $\alpha = P/N$. When α is small, i.e. few object manifolds are in a high ambient dimension, it's easy to find a separating hyperplane for a random dichotomy[1] of the manifolds. When α is large, too many categories are squeezed in a small ambient dimension, rendering the representations highly inseparable. *Manifold capacity* refers to the critical load, $\alpha_C = P/N$, defined by the critical number of object manifolds, P, that can be linearly separated given N features. Above α_C, most dichotomies are inseparable, and below α_C, most are separable [23;24]. This framework generalizes the notion of the perceptron storage capacity [25], re-defining the unit for counting capacity as 'object manifolds' rather than individual points. The manifold capacity thus serves as a measure of the linearly decodable information about object identity per unit, and it can be measured from data in two ways:

Empirical Manifold Capacity, α_{SIM}: the manifold capacity can be measured empirically with a bisection search to find the critical number of features N such that the fraction of linearly separable random dichotomies is close to $1/2$.

Mean Field Theoretic Manifold Capacity, α_{MFT}: can be estimated using the replica mean field formalism with the framework introduced by [23;24]. α_{MFT} is estimated from the statistics of *anchor points* (shown in Fig. 1(c)), $\tilde{s}$, a representative point for a linear classification[2].

The manifold capacity for point-cloud manifolds is lower bounded by $\alpha_{LB} = P/N = 2/M$ due to Cover's theorem [23]. In this work, we show α_{MFT}/α_{LB} for a comparison between datasets with different lower bounds.

Manifold capacity is closely related to the underlying geometric properties of the object manifolds. Recent work demonstrates that the manifold classification capacity can be predicted by an object manifold's *Manifold Dimension*, D_M, *Manifold Radius*, R_M, and the correlations between the centroids of the manifolds [22;23;24]. These geometrical properties capture the statistical properties of the anchor points, the representative support vectors of each manifold relevant for the linear classification, which change as the choice of other manifolds vary.

Manifold Dimension, D_M: D_M captures the dimensions realized by the anchor point from the guiding Gaussian vectors shown in Fig. 1(c), and estimates the average embedding dimension of the manifold contributing to the classification. This is upper bounded by $\min(M, N)$, where M is the number of points per each manifold, and N, the feature dimension. In this work, $M < N$, and we present D_M/M for fair comparison between different datasets.

Manifold Radius, R_M: R_M is the average distance between the manifold center and the anchor points as shown in Fig. 1(c). Note that R_M is the size relative to the norm of the manifold center, reflecting the fact that the relative scale of the manifold compared to the overall distribution is what matters for linear separability, rather than the absolute scale.

Center Correlations, ρ_{center}: ρ_{center} is another geometric property capturing how correlated the locations of these object manifolds are, and is measured as the average of pairwise correlations between object manifold centers [24].

It has been suggested that the capacity is inversely correlated with D_M, R_M, and center correlation [23;24]. Details for computing anchor points can be found in the description of the MFT-based algorithm (Appendix).

[1] Here, we define a random dichotomy as an assignment of random ± 1 labels to each manifold

[2] See Appendix for exact relationship between $\tilde{s}$ and capacity, the outline of the code, and a demonstration that MFT manifold capacity matches the empirical capacity (given in Fig. A1)

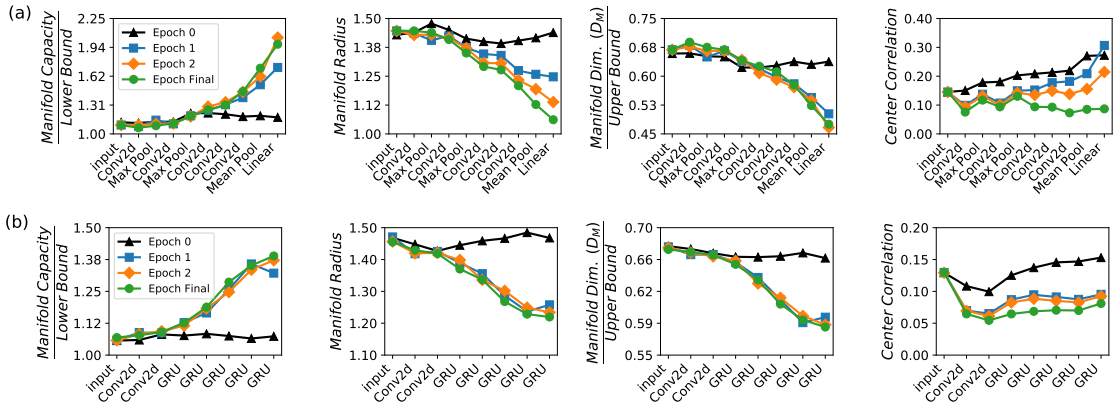

Figure 2: **Word manifold capacity emerges in both (a) the CNN word classification model, and (b) the end to end ASR model (DS2).** (a): As expected, CNN model trained with explicit word supervision (blue lines) exhibits strong capacity in later layers, compared to the initial weights (black lines). This increase is due to reduced radius and dimension, as well as decorrelation. (b): A similar trend emerges in DS2 without training with explicit word supervision. In both, capacity is normalized against the theoretical lower bound (See Methods).

2.2 Models and datasets

We examined two speech recognition models. The first model is a CNN model based on the architecture in[13], trained on the word recognition task (full architecture can be found in Table 2). We trained on two second segments from a combination of the WSJ Corpus[28] and Spoken Wikipedia Corpus[29], with noise augmentation from audio set backgrounds[30]. For more training and performance details, please see the Appendix . Word manifolds from the CNN dataset were measured using data from the WSJ corpus. Each of the $P = 50$ word manifolds consist of $M = 50$ speakers saying the word.

The second is an end-to-end ASR model, Deep Speech 2 (DS2)[4], based on an open source implementation[3]. DS2 is trained to produce accurate character-level output with the Connectionist Temporal Classification (CTC) loss function[31]. The full architecture can be found in Table 1 in Appendix, along with performance details. Word manifolds were taken from the test portion of the LibriSpeech dataset [4]. $P = 50$ words with $M = 20$ examples each were selected, ensuring each example came from a different speaker.

For each layer of the CNN and DS2 models, the activations were measured for each exemplar and 5000 random projections with unit length were computed on which to measure the geometric properties. For temporal analysis in the RNN(DS2) model, full features were extracted for each time step.

3 Results

3.1 Untangling of words across deep network layers and training epochs

We first investigated the CNN model, which was trained to identify words from a fixed vocabulary using the dataset described in 2.2. Since this model had explicit word level supervision, we observed that the words become more separable (higher capacity) in deeper network layers (Figure 2a) as expected. The emergence of word manifold untangling was not observed with the initial weights of the model (black lines). Furthermore, MFT metrics reveal that this increased word capacity in later layers is due to both a reduction in the manifold radius and the manifold dimension (Figure 2a).

End-to-end ASR systems are not trained to explicitly classify words, owing to the difficulty in annotating large datasets with word level alignment, and the large vocabulary size of natural speech. Instead, models such a Deep Speech 2 are trained to output character-level sequences. Despite not being trained with word labels, the word level untangling was observed on the Librispeech dataset (Figure 2b).

Surprisingly, across the CNN and recurrent DS2 architectures, the trend in the manifold metrics were similar. The manifold capacities improve in downstream layers, and the reduction in manifold dimension and radius similarly occurs

[3]https://github.com/SeanNaren/deepspeech.pytorch

[4]See the Appendix for more details on the construction and composition of the word dataset used for experiments on this model.

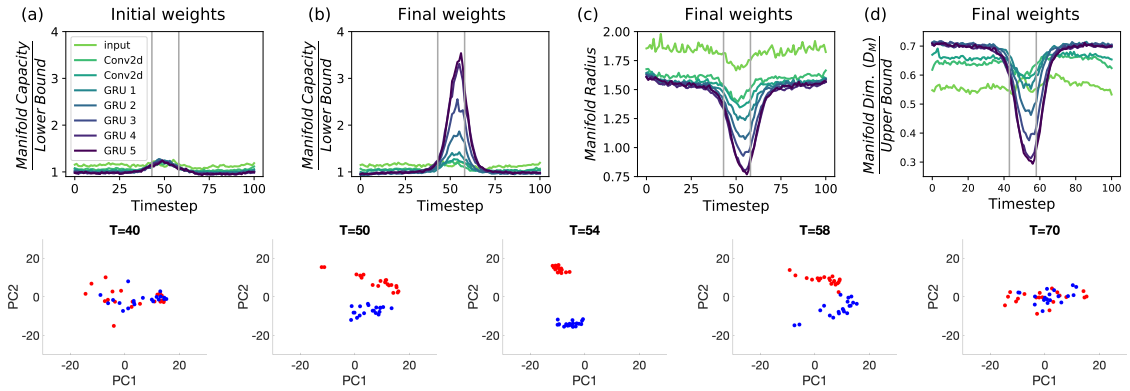

Figure 3: **Untangling of word manifolds in input timesteps.** (Top) Evolution of Librispeech word manifolds in timesteps, RNN (DS2) model (hypothesized in Fig 1). (a) Epoch 0 model, capacity (b-d) fully trained model, (b) capacity, (c) manifold radius, (d) manifold dimension. Vertical lines show the average word boundaries. (Bottom) Untanglement of two words over timesteps (T=40 to 70) in GRU 5 layer of DS2, projected to 2 PCs.

in downstream layers. Interestingly, word manifolds increases dramatically in the last layer of CNN, but only modestly in the last layers of DS2, perhaps owing to the fact that CNN model here is explicitly trained on word labels, while in the DS2, the word manifolds are emergent properties. Notably, the random weights of the initial model increase correlation across the layers in both networks, but the training significantly decreases center correlation in both models. The results in the early stages of training in CNN and DS2 indicate that the capacity, manifold dimension, manifold radius, and center correlations quickly converge to those measured on the final epochs.

3.2 Untangling of words over time

Here, we compute these measures of untangling on each time step separately. This approach reveals the role of time in the computation, especially in recurrent models processing arbitrary length inputs.

Figure 3 shows the behavior of capacity, manifold radius, and manifold dimension over the different time steps in the recurrent layers of the end-to-end ASR model (DS2) for the word inputs used in Sec. 3.1. As is perhaps expected, the separability is at the theoretical lower bound for times far away from the word of interest, and peaks near the location of the word. This behavior arises due to the decrease in radius and dimension.

The capacity measured at each input timesteps has a remarkable peak relative capacity of 3.6 (Fig. 3), much larger than that of the random projection features, at 1.4 (Fig. 2). This suggests the sequential processing massages the representation in a meaningful way, such that a snapshot at a peak timestep has a well separated, compressed representation, captured by the small value of D_M and R_M. Analogous to 1(d), the efficiency of temporal separation is illustrated in Fig. 3, bottom.

4 Conclusion

In this paper we studied the emergent geometric properties of speech objects and their linear separability, measured by manifold capacity. Across different networks and datasets, we find that linear separability of auditory class objects improves across the deep network layers, consistent with the untangling hypothesis in vision[1]. Word manifold's capacity arises across the deep layers, due to emergent geometric properties, reducing manifold dimension, radius and center correlations. Characterization of manifolds across training epochs suggests that word untangling is a result of training, as random weights do not untangle word information in the CNN or DS2. As ASR systems have representations evolve on the timescale of input sequences, we find that separation between different words emerges temporally, by measuring capacity and geometric properties at every frame.

Our methodology and results suggest many interesting future directions. We hope that our work will motivate: (1) the theory-driven geometric analysis of representation untangling in tasks with temporal structure; (2) the search for the mechanistic relation between the network architecture, learned parameters, and structure of the stimuli via the lens of geometry.

References

[1] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[2] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

[3] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

[4] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[5] Yonatan Belinkov and James Glass. Analyzing hidden representations in end-to-end automatic speech recognition systems. In *Advances in Neural Information Processing Systems*, pages 2441–2451, 2017.

[6] Yonatan Belinkov. *On internal language representations in deep learning: An analysis of machine translation and speech recognition*. PhD thesis, Massachusetts Institute of Technology, 2018.

[7] Tasha Nagamine, Michael L Seltzer, and Nima Mesgarani. Exploring how deep neural networks form phonemic categories. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Tasha Nagamine, Michael L Seltzer, and Nima Mesgarani. On the role of nonlinear transformations in deep neural network acoustic models. In *Interspeech*, pages 803–807, 2016.

[9] Yu-Hsuan Wang, Cheng-Tao Chung, and Hung-yi Lee. Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries. *Interspeech*, 2017.

[10] Shuai Wang, Yanmin Qian, and Kai Yu. What does the speaker embedding encode? In *Interspeech*, pages 1497–1501, 2017.

[11] Zied Elloumi, Laurent Besacier, Olivier Galibert, and Benjamin Lecouteux. Analyzing learned representations of a deep asr performance prediction model. *arXiv preprint arXiv:1808.08573*, 2018.

[12] Andreas Krug, René Knaebel, and Sebastian Stober. Neuron activation profiles for interpreting convolutional speech recognition models. 2018.

[13] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[14] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017.

[15] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John E Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *International Conference on Learning Representations*, 2016.

[16] Peiran Gao, Eric Trautmann, M Yu Byron, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, page 214262, 2017.

[17] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[18] David GT Barrett, Ari S Morcos, and Jakob H Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology*, 55:55–64, 2019.

[19] Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. Perceptual straightening of natural videos. *Nature neuroscience*, page 1, 2019.

[20] Olivier J Hénaff and Eero P Simoncelli. Geodesics of learned representations. 2016.

[21] Raja Giryes, Guillermo Sapiro, and Alexander M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Trans. Signal Processing*, 64(13):3444–3457, 2016.

[22] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Linear readout of object manifolds. *Physical Review E*, 93(6):060301, 2016.

[23] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

[24] Uri Cohen, SueYeon Chung, Daniel D. Lee Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *bioRxiv*, page 10.1101/644658, 2019.

[25] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

[26] HS Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[27] Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03014, 2013.

[28] Douglas B Paul and Janet M Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

[29] Arne Köhn, Florian Stegen, and Timo Baumann. Mining the spoken wikipedia for speech data and beyond. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[30] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[31] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Appendices

A Details on measuring empirical and theoretical manifold capacity

A.1 Empirical Manifold Capacity

Here we provide a detailed description for empirically finding a manifold capacity, for a given number of object class manifolds , P, by finding a critical number of feature dimensions, N_c, such that the fraction of separable dichotomies of random assignment of +/- labels to a given manifolds is at 0.5 on average (Fig. A1). If the feature dimension N is larger than critical N_c, the fraction of separable dichotomies will be close to 1 (hence, the system is in a linearly separable regime, Fig. A1). If the feature dimension N is smaller than critical N_c, the fraction of separable dichotomies will be close to 0 (the system is in a linearly in-separable regime, Fig. A1). The algorithm finds N_c by doing a bisection search on N, such that "fraction of linearly separable dichotomies" for N_c is 0.5, midpoint between 1 (separable) and 0 (inseparable) on average. At the cricical N_c, the capacity is defined to be P/N_c. In our experiments shown in Fig. A1, we used randomly sampled 101 dichotomies, to compute fraction of linear separability.

A.2 Mean-Field Theoretic (MFT) Manifold Capacity and Geometry

Here, we provide a summmary for finding a theoretical estimation of manifold capacity using mean-field theoretic approach. It has been proven that the general form of the inverse MFT capacity, exact in the thermodynamic limit, is given by:

$$\alpha_{MFT}^{-1} = \left\langle \frac{\left[t_0 + \vec{t} \cdot \tilde{s}(\vec{t})\right]_+^2}{1 + \left\|\tilde{s}(\vec{t})\right\|^2} \right\rangle_{\vec{t},t_0}$$

where $\langle \ldots \rangle_{\vec{t},t_0}$ is an average over random D- and 1- dimensional vectors $\vec{t}, t_0$ whose components are i.i.d. normally distributed $t_i \sim \mathcal{N}(0,1)$.

Central to this framework is the notion of *anchor points*, $\tilde{s}$ (section 2.1 in the main text), uniquely given by each $\vec{t}, t_0$, representing contributions from all other object manifolds, in their random orientations. For each $\vec{t}, t_0$, $\tilde{s}$ is uniquely defined as a subgradient that obeys the KKT conditions, hence, $\tilde{s}$ in KKT interpretation, represents a weighted sum of support vectors contributing to the linearly separating solution.

These anchor points play a key role in estimating the manifold's geometric properties, given as: $R_M^2 = \left\langle \left\|\tilde{s}(\vec{T})\right\|^2 \right\rangle_{\vec{T}}$ and $D_M = \left\langle \left(\vec{t} \cdot \hat{s}(\vec{T})\right)^2 \right\rangle_{\vec{T}}$ where $\hat{s}$ is a unit vector in the direction of $\tilde{s}$, and $\vec{T} = (\vec{t}, t_0)$, which is a combined coordinate for manifold's embedded space, and manifold's center direction (in general, if we compute the geometric properties in the ambient dimension, it includes both the embedded space and center direction).

The manifold dimension measures the dimensional spread between $\vec{t}$ and its unique anchor point $\tilde{s}$ in D dimensions (the coordinates in which each manifold is embedded).

In high dimension, the geometric properties predict the MFT manifold capacity, by

$$\alpha_{MFT} \approx \alpha_{Ball}\left(R_M,\, D_M\right) \tag{1}$$

where,

$$\alpha_{Ball}^{-1}(R, D) = \int_{-\infty}^{R\sqrt{D}} Dt_0 \frac{(R\sqrt{D} - t_0)^2}{R^2 + 1} \tag{2}$$

Note that the above formalism is assuming that the manifolds are in random locations and orientations, and in real data, the manifolds have various correlations. So, we apply the above formalism onto the data that has been projected into the null spaces of centers, using the method proposed by [24].

The validity of this method is shown in Fig. A1, where we demonstrate the good match between the empirical manifold capacity (computed using a method in Section. A.1) and the mean-field theoretical estimation of manifold capacity (using the algorithm provided in this section).

For more details on the theoretical derivations and interpretations for the mean-field theoretic algorithm, see [24] [23].

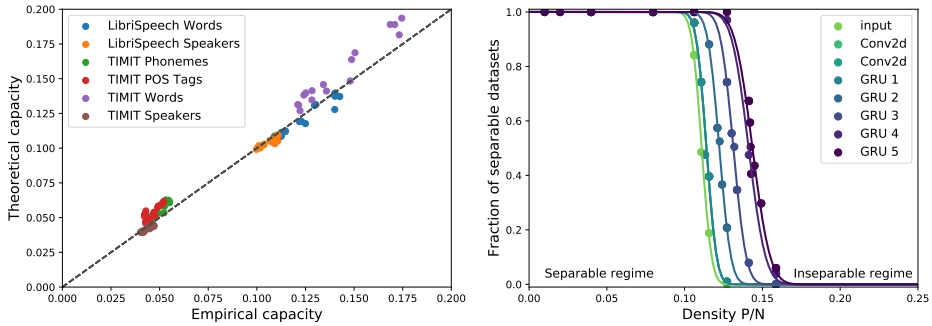

Figure A1: **Measured capacity matches theoretical prediction.** (a) Across multiple datasets (TIMIT, Librispeech) and manifolds (words, speakers, phonemes, part-of-speech tags) for the DS2 model, the measured capacity closely matches the theoretical capacity. Dotted line indicates unity. (b) Empirical capacity is measured by a bisection search for critical N s.t. fraction of separable datasets cross 0.5

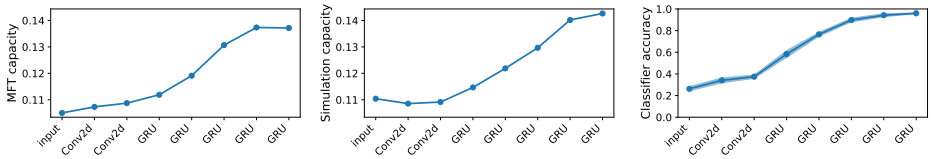

Figure A2: Measured capacity matches theoretical prediction, and trends are also reflected in the generalization error of a linear classifier in the LibriSpeech word manifolds experiment

Algorithm 1 compute_geometric_properties: Mean-field theoretic capacity and geometry for data

Function compute_geometric_properties($\{X^\mu\}$)

Input: Categorical data $\left\{X_i^\mu \in \mathbb{R}^N\right\}_{i\in[1..M_\mu]}^{\mu=1..P}$ (P=#Manifolds, M_μ=#Samples per μth Manifold)

1. Subtract global mean and update $\left\{X_i^\mu \in \mathbb{R}^N\right\}_{i\in[1..M_\mu]}^{\mu=1..P}$
2. Compute centers of each manifold $\{\vec{c}^\mu \in \mathbb{R}^N\}^{\mu=1,\ldots,P}$
3. Compute center correlations δ_{CC}
4. Find subspace shared by manifold centers* : $\{C^\mu\} = \text{find\_center\_subspace}\left(\{X^\mu\}\right)$
5. Project original data into null space of center subspaces*: $\left\{X^{\perp\mu}\right\} = \text{find\_residual\_data}\left(\{X^\mu, C^\mu\}\right)$
6. Normalize data s.t. center norms are 1**: $X^{0\perp\mu} = \text{manifold\_normalize}\left(X^{\perp\mu}\right)$
7. For $\mu = 1..P$, calculate geometry**: $D_M^\mu, R_M^\mu, \alpha_c^\mu = \text{manifold\_geometry}\left(X^{0\perp\mu}\right)$

Output: $\{D_M^\mu\}_{\mu=1}^P, \{R_M^\mu\}_{\mu=1}^P, \{\alpha_c^\mu\}_{\mu=1}^P$

* is based on [24], and ** is based on [23].

B Verification and control experiments

B.1 Theory matches empirical capacity

To validate the MFT capacity measure on these datasets, we carried out additional experiments comparing the results from the bisection search to find the empirical capacity α_{SIM} and the result obtained from the MFT calculation, α_{MFT} on each of the datasets used in the experiments on the ASR model. Figure A1 (Left) shows the agreement between the two measures, while (Right) shows the transition between the separable and inseparable regimes found during the search for α_{SIM}.

B.2 Trends in linear classifier accuracy

As further verification of the MFT analysis, we also compared the trends in capacity, both theoretical α_{MFT} and empirical α_{SIM} to trends observed in the generalization performance of a linear classifier. Figures A2 shows the comparison to the generalization

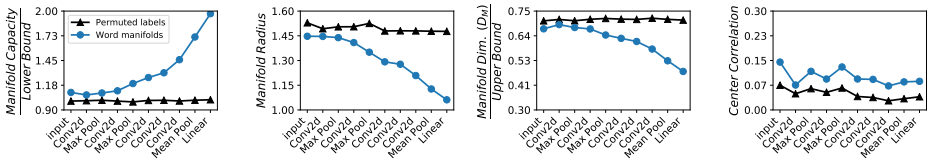

Figure A3: Manifold untangling dissapears when class labels are permuted, CNN word manifolds on word trained network

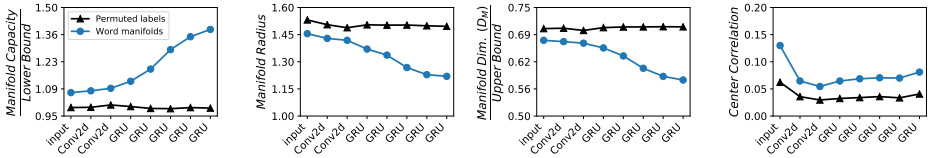

Figure A4: Manifold untangling dissapears when class labels are permuted, Librispeech words experiment

accuracy of a one-vs-rest logistic regression classifier trained on 80% of the manifold data. Here, the classifier was trained 10 random train/tests splits of the manifold data, and the average over manifolds and trials is reported here along with the standard deviation of the mean. We find that the trends observed in the the classifier performance follow those seen in the measures of capacity.

B.3 Experiments with permuted labels

The trends observed in the MFT analysis should vanish when the activations are not grouped by class label. In Figures A3 and A4 we verify that the observed trends do not occur when the class labels are randomized.

C DS2 Model details

The ASR model used in experiments is based on the popular Deep Speech 2 architecture[4]. A complete specification of the model is given in Table 1.

Table 1: End-to-end ASR model architecture

Layer	Type	Size
0	Input	$T \times 161$ spectral features
1	2D Convolution	32 filters of shape 41×11, stride 2
2	2D BatchNorm	-
3	HardTanh	-
4	2D Convolution	32 filters of shape 21×11, stride 2 in time only
5	2D BatchNorm	-
6	HardTanh	-
7	Bidirectional GRU	800
8	1D BatchNorm[4]	-
9	Bidirectional GRU	800
10	1D BatchNorm[4]	-
11	Bidirectional GRU	800
12	1D BatchNorm[4]	-
13	Bidirectional GRU	800
14	1D BatchNorm[4]	-
15	Bidirectional GRU	800
16	1D BatchNorm[4]	-
17	Linear	800×29

This model was trained on the 960 hour training portion of the LibriSpeech dataset[32] for 68 epochs with an initial learning rate of 0.0003 and a learning rate annealing of 1.1. The trained model has a word error rate (WER) of 12%, 22.7% respectively on the clean and other partitions of the test set without the use of a language model. The WER for different training epochs is shown in Figure A5. The model also performs reasonably well on the TIMIT dataset, with a WER of 29.9% without using a language model.

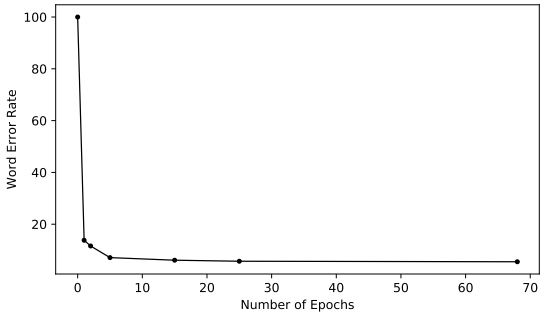

Figure A5: **LibriSpeech Word Error Rate (WER) on the test set for different epochs** similarly to the results in capacity, the performance quickly saturates after a few epochs.

D CNN Model training procedures

Table 2: Word (and Speaker) CNN Model Architecture, same as /citekell2018task but with batch normalization instead of local response normalization.

Layer	Type	Size
0	Input	256×256 cochleagram
1	2D Convolution	96 filters of shape 9×9, stride 3
2	ReLu	-
3	MaxPool	window 3×3, stride 2
4	2D BatchNorm	-
5	2D Convolution	256 filters of shape 5×5, stride 2
6	ReLu	-
7	MaxPool	window 3×3, stride 2
8	2D BatchNorm	-
9	2D Convolution	512 filters of shape 3×3, stride 1
10	ReLu	-
11	2D Convolution	1024 filters of shape 3×3, stride 1
12	ReLu	-
13	2D Convolution	512 filters of shape 3×3, stride 1
14	ReLu	-
15	AveragePool	window 3×3, stride 2
16	Linear	4096 units
17	ReLu	-
18	Dropout	0.5 prob during training
19	Linear	Num Classes

For word recognition, we trained on two second segments from a combination of the WSJ Corpus [28] and Spoken Wikipedia Corpus [29], with noise augmentation from audio set backgrounds [30]. We selected two second sound segments such that a single word occurs at one second. For the training set, we selected words and speaker classes such that each class contained 50 unique cross class labels (ie 50 unique speakers had to say each of the word classes). We also selected words and speaker classes that each contained at least 200 unique utterances, and ensured that each category could contain a maximum of 25% of a single cross category label (ie for a given word class, a maximum of 25% of the utterances could come from a single speaker), the maximum number of utterances in any word category was less than 2000, and the maximum number of utterances within any speaker category was less than 2000. Data augmentation during training consisted jittering the input in time and placing the exemplars on different audioset backgrounds.

The resulting training dataset contained 230,357 segments in 433 speaker classes and 793 word classes. The word recognition network achieved a WER of 22.7% on the test set, and the speaker recognition network achieved an error rate of 1% on the test set.

E Extra plots

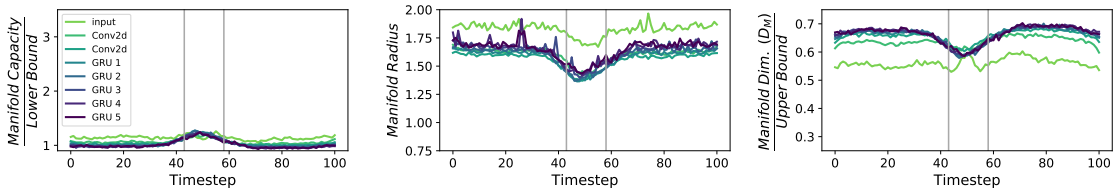

Figure A6: **Untangling of word manifolds in input time steps before training DS2.** No improvement over the layers, and slightly more information in the midpoint.