We are grateful to the reviewers for their insightful and constructive comments. We appreciate that each reviewer found our approach to be a useful step towards a better understanding of ASR systems. Below we respond to the main questions and concerns, paraphrased for brevity.

To All Reviewers: To improve accessibility of the method, we will open source the analysis code, and clarify our procedures in the revision.

Reviewer 1 (R1): *Random weight baseline*: We highlight a few pieces of evidence that suggest the differences in geometry are due to learning rather than the changing weight norms: the effects of training vanish in all cases when the manifold labels are permuted (SM Fig. 14-20), opposing trends are seen in when measuring word or speaker manifolds (Fig. 2 and Fig. 4), and these trends change based on the task the model is trained on (Fig. 4 top).

Error bars: In the present article, we used a single typical training run of two popular open-sourced models (word-CNN and DS2). This is consistent with the methods of Refs. [6, 7, 25]. We demonstrate the robustness across different random projections of network features and initializations for the mean-field metric calculations. An example of 95% confidence intervals calculated in this way is included for word-CNN (Fig A1(a)) and we will update the remaining figures in the document similarly.

Reviewer 2 (R2): *Training dataset differences for the CNN and the DS2 models*: The discrepancy in datasets is due to the necessity of word-aligned speech for training the word-CNN, which is not publicly available for LibriSpeech. The "CNN dataset" is adapted from that used in Ref. [14], which we supplement with words from Spoken Wikipedia Corpus (SWC) to diversify the word instances and provide more balanced speaker classes for the speaker trained model. The details of this training dataset construction are explained in SM section 3, and in the revision we will improve the clarity in the main text when referencing these datasets and models.

Connection between phoneme and character manifolds: We agree that our method could be useful to illuminate the complex relation between characters and phones in ASR systems. Here we share a preliminary result in Fig A1(b-d), showing that character-level manifolds also emerge across the layers of DS2. The relative increase in the manifold capacity in the last recurrent layer of DS2 compared to the input layer is slightly larger in characters than phonemes, consistent with recent findings by Belinkov, Ahmed, and Glass, arXiv:1907.04224 (2019). The further exploration into what this method can tell us about the grapheme to phoneme relationship is a promising direction for future work but would be better served as a separate paper.

Reviewer 3 (R3): *Clarification about the scope of our experiments and findings*: While it is true that some of our results are consistent with current beliefs, our work provides empirical evidence for these beliefs, and goes deeper by connecting object information to underlying feature geometry as enabled by the theory. The presented results also include two different ASR models with different motifs (convolutional and recurrent), one trained with word label units and the other with character sequences, and we see similar behavior in both experimental settings.

Clarifying ambient dimension and capacity's relation to vocabulary size: Ambient dimension N is the feature dimension. If the manifold capacity (P/N_c) is large given fixed P (size of vocabulary), then the required feature dimension for separability, N_c, is small, and the classes are linearly separable (they are untangled), as long as $N > N_c$. More insights on manifold capacity can be found in Ref [24, 25].

Correspondence of capacity and Word Error Rate (WER): R3 raises an interesting question of the direct relationship between the capacity measures and WER. To address this, we share an additional result on the relationship between capacity and WER in DS2 over different epochs of the training (Fig A1(e)) showing that as capacity increases, WER decreases. We also note that the network-level correspondence between the manifold capacity and the accuracy is shown in the prior work in vision (Ref. [25]).

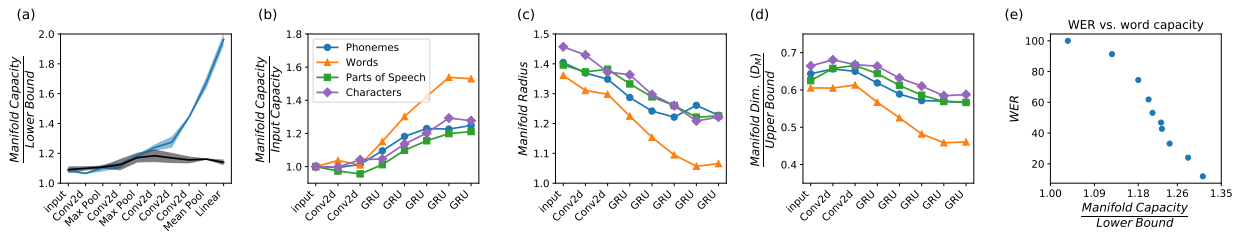

Figure A1: (a) word manifold capacity in word-CNN, (blue) after training, (black) before training, (b-d) character manifolds compared with phonemes, words, POS manifolds, (e) WER vs word capacity on the test set