CNNs that robustly compute vowel harmony do not explicitly represent phonological tiers

Jane Li and Alan Zhou
Department of Cognitive Science,
Johns Hopkins University
Baltimore, MD
{sli213, azhou23}@jhu.edu

1 Introduction

Linguistic and model-theoretic analyses of long-distance phonology postulate the existence of phonological tiers (Goldsmith, 1976; Heinz et al., 2011). For example, vowel harmony may be analyzed as a process that projects vowels (but not consonants) onto a tier and ensures that all sounds on the tier have the same feature (e.g., [±front] in Turkish vowel harmony, Clements et al. (1982)).

Li and Zhou (under review) recently demonstrated that convolutional neural networks (CNNs) learning a toy example of vowel harmony (§2) on short strings robustly generalize the pattern to much longer strings. One explanation is that these CNNs have independently recovered an "algorithm" that is consistent with the tier projection analysis. Alternatively, these models may have uncovered an approximation of this system, or an entirely different system that robustly generalizes to long lengths. This work investigates these hypotheses via various interpretability methods. In particular, we search for evidence for a "strong" implementation of tier projection, in which these CNNs exactly implement the tier-projection and feature-matching analyses described above.

2 Model and toy language

We follow the architecture of the CNN string recognizer described in [4]. Strings are passed as a block of one-hot character encodings into a convolutional neural network consisting of 4 one-dimensional layers. The output of this CNN is passed through a global max-pool, followed by a single fully connected layer that outputs for each string a binary classification score between 0 and 1. Strings with score above 0.5 are treated as belonging to the recognizer's string language (e.g. the set of strings obeying an unbounded vowel harmony rule). Each convolutional layer is parameterized with a kernel size of 3, a channel size of 256, and a stride of 1 with same padding.

CNNs were trained on an artificial string acceptance task designed to emulate a pattern of transparent unbounded vowel harmony. Artificial strings are sampled by generating syllables roughly obeying the sonority sequencing principle with a vowel inventory {a, e, o, u, ä, ë, ö, ü}, with the constraint of vowels agreeing in the presence of trema (V̈) or absence of trema (V) in harmonious strings. Models learned to recognize if a given string obeys the vowel harmony rule, obtaining perfect test accuracy even over strings much longer than those seen during training.

3 CNNs do not implement exact tier projection

We first investigate the hypothesis that these CNN models are explicitly performing "hard" tier projection. That is, there exists some intermediate layer of the CNN in which vowels (but not consonants) are being projected. If this is the case, we hypothesize that unprojected consonant strings such as [spl] and [spr] should not be distinguishable from one another in terms of activation at that layer. We tested this prediction by decoding the consonants [l] from [r] and the voiceless stops [p,t,k] from each other. For each set of sounds, we selected all attested length-3 consonant clusters where sounds in the set can appear interchangeably. We obtained activations for all of these clusters and decoded the presence of one target sound in the sound set (e.g., [spr] has [p], but [str] and [skr] do not). We find that all sound sets are reliably decodable (Fig. 1A).

Although the performance of the decoder drops off towards later layers, it remains substantially higher than that of a conservative baseline. We observe a similar trend when we attempt to decode sound presence in CVCVC sequences (e.g., is [p] present in [palar] vs. [torel]?). However, we note that while decoding accuracy falls off in later layers for all sounds, consonants consistently fall off

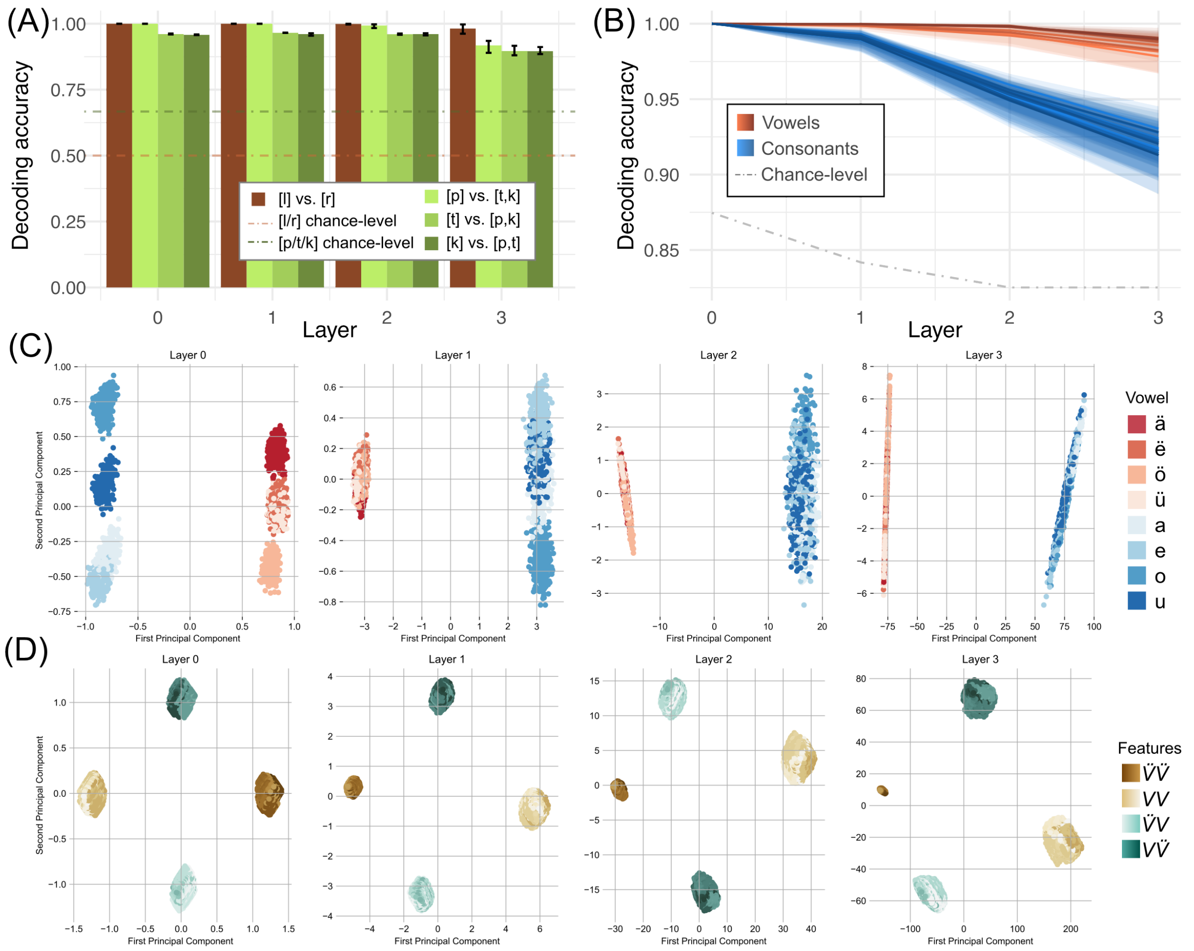

Figure 1: **A.** Decoding accuracy (via ridge regression) for [l] against [r] and [p, t, k] against each other. **B.** Decoding accuracy for presence of individual sounds. Error bars/ribbons in (A, B) indicate range of accuracies across 5 runs. **C and D.** Intermediate activations projected along the first two principal components obtained from PCA over all possible CVC inputs (C) and all possible CVCVC inputs (D). Activations are taken after the application of ReLU in each layer, and flattened for PCA. Each individual set of activations is colored by the identity of the vowel in the CVC sequence (C) and by the sequence of vowel features in the CVCVC sequence (D).

more than vowels (Fig. 1B). While we conclude that these CNNs have not learned to perform tier projection exactly, some prioritization for vowels over consonants is observed.

4 CNNs demonstrate feature-based abstraction over vowels

We now turn to ask whether among the vowels, some abstraction has formed to facilitate the computation of vowel harmony, such as that of V̈ as a category vs. V. We investigate this by applying principle component analysis (PCA) to the activations of each convolutional layer in response to all possible CVC sequences (Fig. 1C), and separately for all CVCVC sequences (Fig. 1D). Applying PCA to the CVC inputs, we find evidence that CNN representations do reflect abstract vowel features, with the V-V̈ distinction being strongly captured by the first principal component (PC) in all layers of the network. Applying PCA to the CVCVC outputs yields similar findings, with the first PC capturing the distinction between the two harmonious feature combinations (VV vs. V̈V̈) and the second PC capturing the distinction between the two disharmonious feature combinations (VV̈ vs. V̈V). We do note, however, that neither of these dimensions seem to reflect the distinction between harmonious and disharmonious feature sequences itself. Preliminary examination suggests that this distinction may be found in the third principal component, though perhaps in a less robust manner than the distinctions described above.

5 Discussion

5.1 A soft implementation of tiers

Altogether the results indicate that the trained CNNs are not implementing an algorithm that fully resembles strict tier projection. However, results do point toward a soft implementation of tiers. Under this hypothesis, the concept of tiers still maps onto a layer of the network, but the layer still has capacity (and learns) to represent other contrasts that are irrelevant to the pattern at-hand. In the case of this toy example, we observe vowel representations become progressively abstract across layers (Fig. 1C) and track vowel bigram information (Fig. 1D), but consonants, which are theoretically irrelevant, are still reliably decodable across all layers (Fig. 1A and 1B). The main prediction is that vowels have "privileged" representations (e.g., better signal within-model) over consonantsthat support computations for the task at hand. This is most evident in the decoding results of Fig. 1B, where vowels consistently better decoded than consonants.

5.2 Alternative theories and their implementations

So far, the possible implementations that have been discussed in this work pertain to a specific framework (tier-based analyses of harmony patterns). It could be the case that the CNNs examined in this study are implementing an algorithm that is consistent with other theories of harmony. Some theories, which assume different forms of input (e.g., articulatory accounts of harmony, Gafos (1999)), may render the models incompatible or be considered as an implementation of intermediary representations. That aside, the methods utilized in this work can be generalized to test hypotheses about what theories a model has learned to implement. A phonological theory makes predictions about what instances (e.g., phonological strings) have shared or contrastive representations. Translating these predictions to signals from model read outs, it predicts that contrastive representations to be decodable or occupy a representational subspace.

5.3 Disambiguating between representations of grammaticality and tier-based representations

We found via PCA that the model has learned to linearly represent the distinction between harmonic and disharmonic vowel sequences. Considering that this is theoretically the only contrast that the model needs to learn to distinguish, these findings are currently confounded with a grammaticality (in other words, output True/False oriented) representation and an algorithmic abstraction of vowel sequences. This should become distinguishable when a model is equipped to learn multiple patterns. Eventually, all patterns have to converge to some representation that supports final True/False decisions, but should have different specific, detectable, representational content for each pattern learned.

6 Conclusion

Our results suggest that these CNNs have converged to a robust solution for unbounded vowel harmony, albeit one that is different from the mechanism of explicit tier projection. In particular, we find that vowels and consonants are both highly decodable from intermediate activations, contrary to what is predicted by an exact tier projection account. However, the intermediate activations of the CNN do reflect robust representations of the vowel features over which harmony is computed, with preliminary evidence for representation of the distinction between harmony and disharmony.

Acknowledgments

We thank the reviewers of this abstract for their helpful feedback. This work is supported by a SSHRC Doctoral Fellowship to JL (752-2024-0291).

References

George N Clements, Engin Sezer, et al. 1982. Vowel and consonant disharmony in turkish. *The structure of phonological representations*, 2:213–255.

Adamantios I Gafos. 1999. *The articulatory basis of locality in phonology*. Ph.D. thesis, Johns Hopkins University.

John Anton Goldsmith. 1976. *Autosegmental phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

Jeffrey Heinz, Chetan Rawal, and Herbert G Tanner. 2011. Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language technologies*, pages 58–64.