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

Jane Li and Alan Zhou {sli213, azhou23}@jhu.edu

JOHNS HOPKINS UNIVERSITY

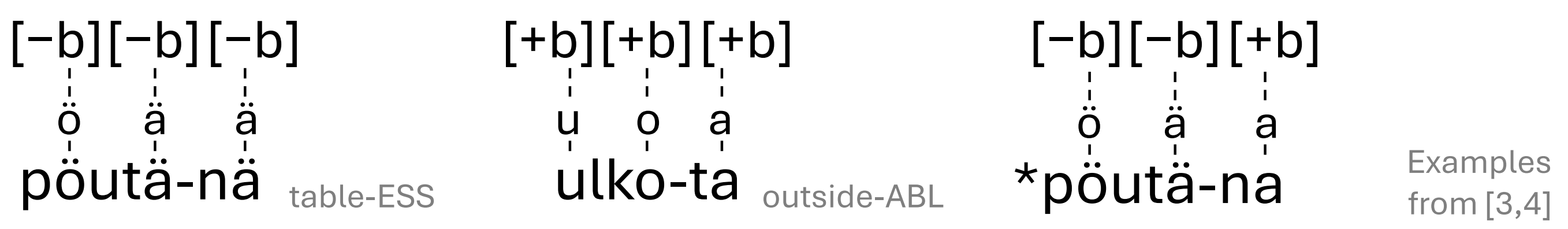
Introduction

In previous work, we found that various **CNNs** easily converge to a solution when trying to learn a toy example of **vowel harmony**.

- Specifically, this solution is highly robust and generalizes to strings far longer than training length.

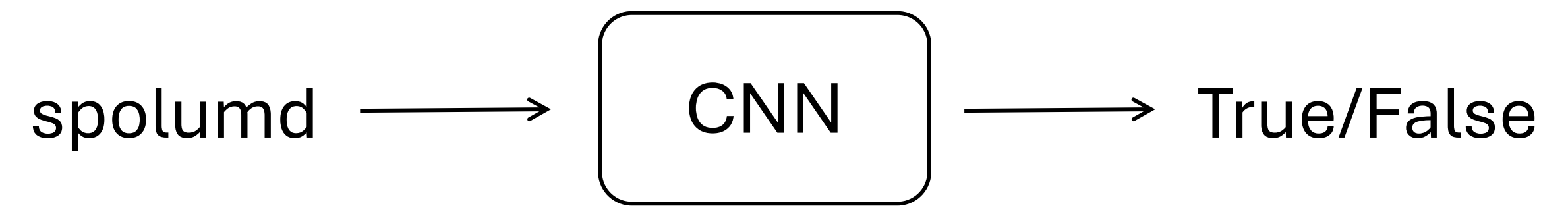
What is the underlying algorithm learned by these CNNs?

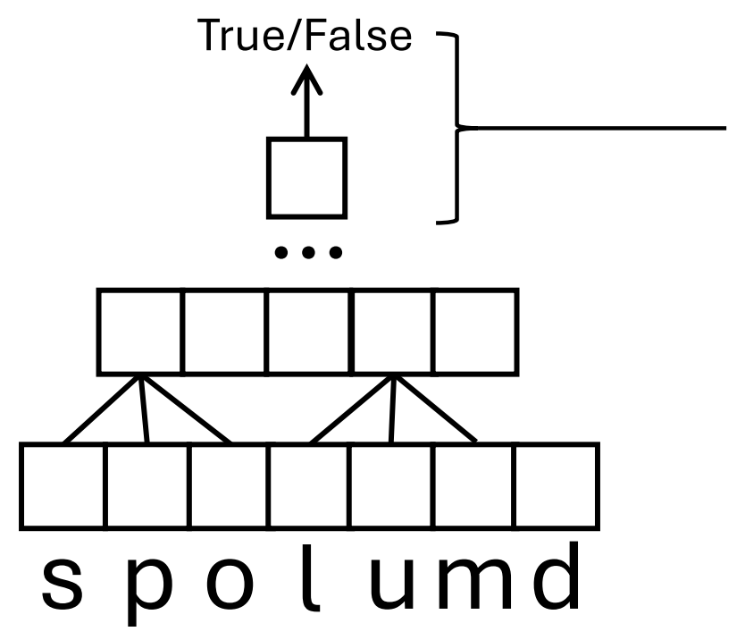
Does the algorithm resemble that of tier-based analyses of harmony patterns? [1,2]



Model & toy language

String recognition: does the string belong to the language?





Model: 4-layer, 128-channel CNN.

Final layer is pooled and put through a linear layer for T/F prediction.

Training for 2-4 σ strings, tested on 5-50 σ strings. Ceiling performance across 5 random seeds.

Toy language: a simplified case of unbounded vowel harmony with simulated phonological strings.

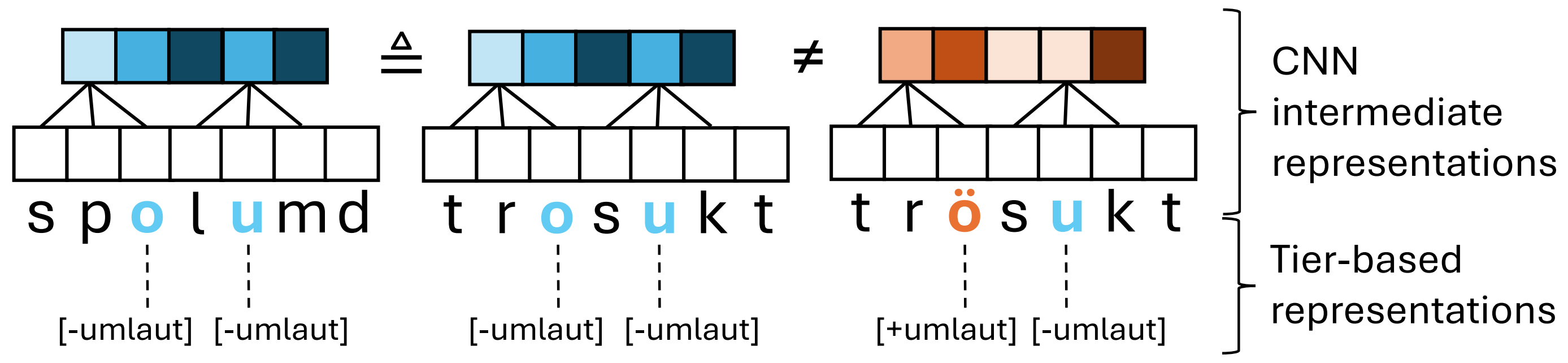
All vowels of the string must have umlaut (äöëü) or be plain (aoeu).

spolumd	spölümd	Accepted (Vs agree)
spolümd	spölumd	Rejected (Vs disagree)

Hypotheses

What algorithm have these models learned to implement such that they promote great length generalization?

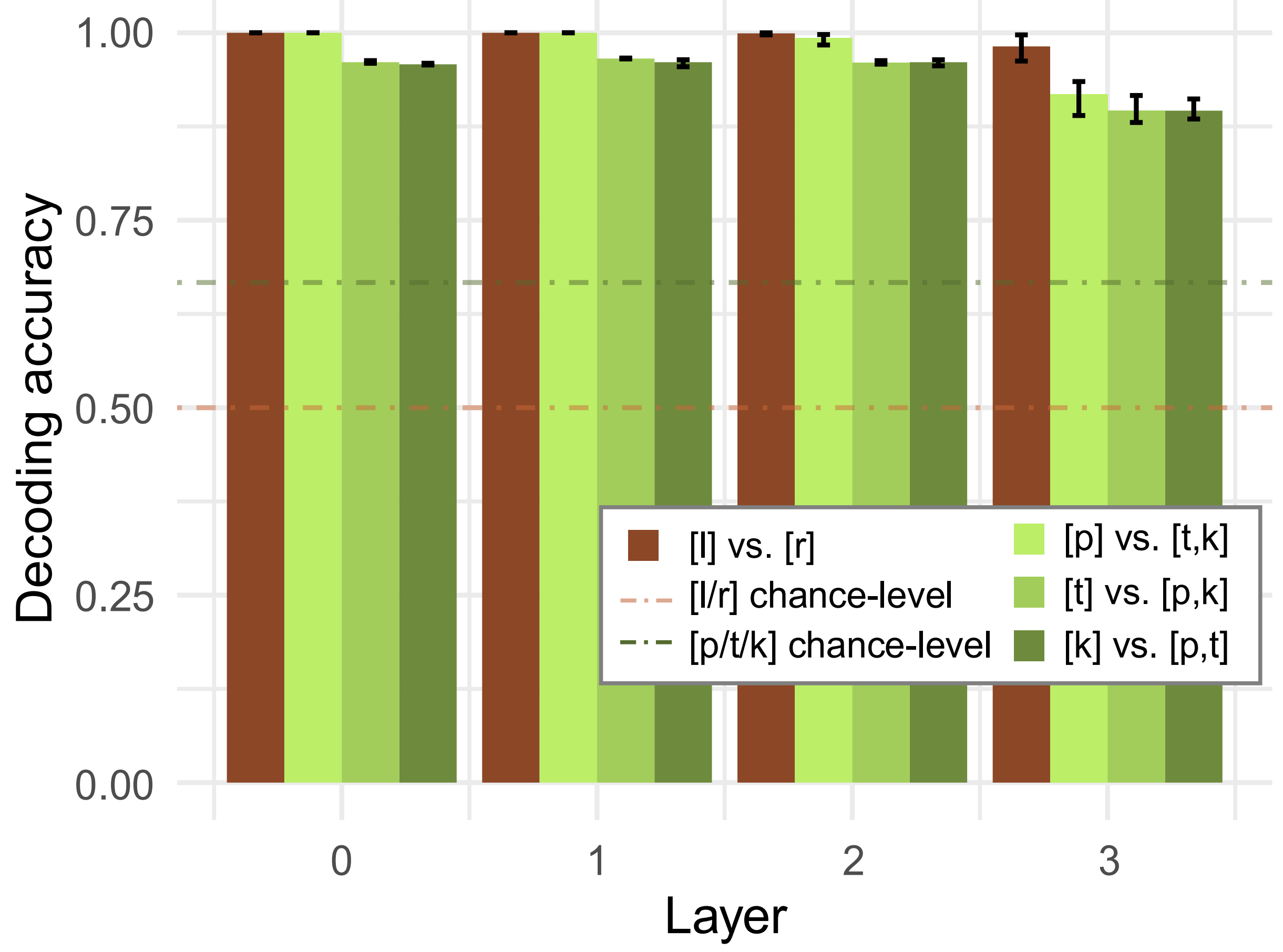
Strong hypothesis: at some layer of the network, only vowels are represented and computed over.



On the lookout for support for a **soft implementation of tiers**:

- Prioritization of representations on the tier over those that are not ⇒ Vs over Cs.
- Abstracted representations at the level of relevant feature/segment group ⇒ Abstraction within umlaut group.

CNNs are not *strictly* implementing tiers



Could CNNs still have properties that allow for approximation of a tier-based account?

Are intermediate representations of strings with different consonants linearly decodable?

Prediction from strong hypothesis: No ⇒ decoding at chance level.

Linear classifier fit over activations of minimal pair strings:

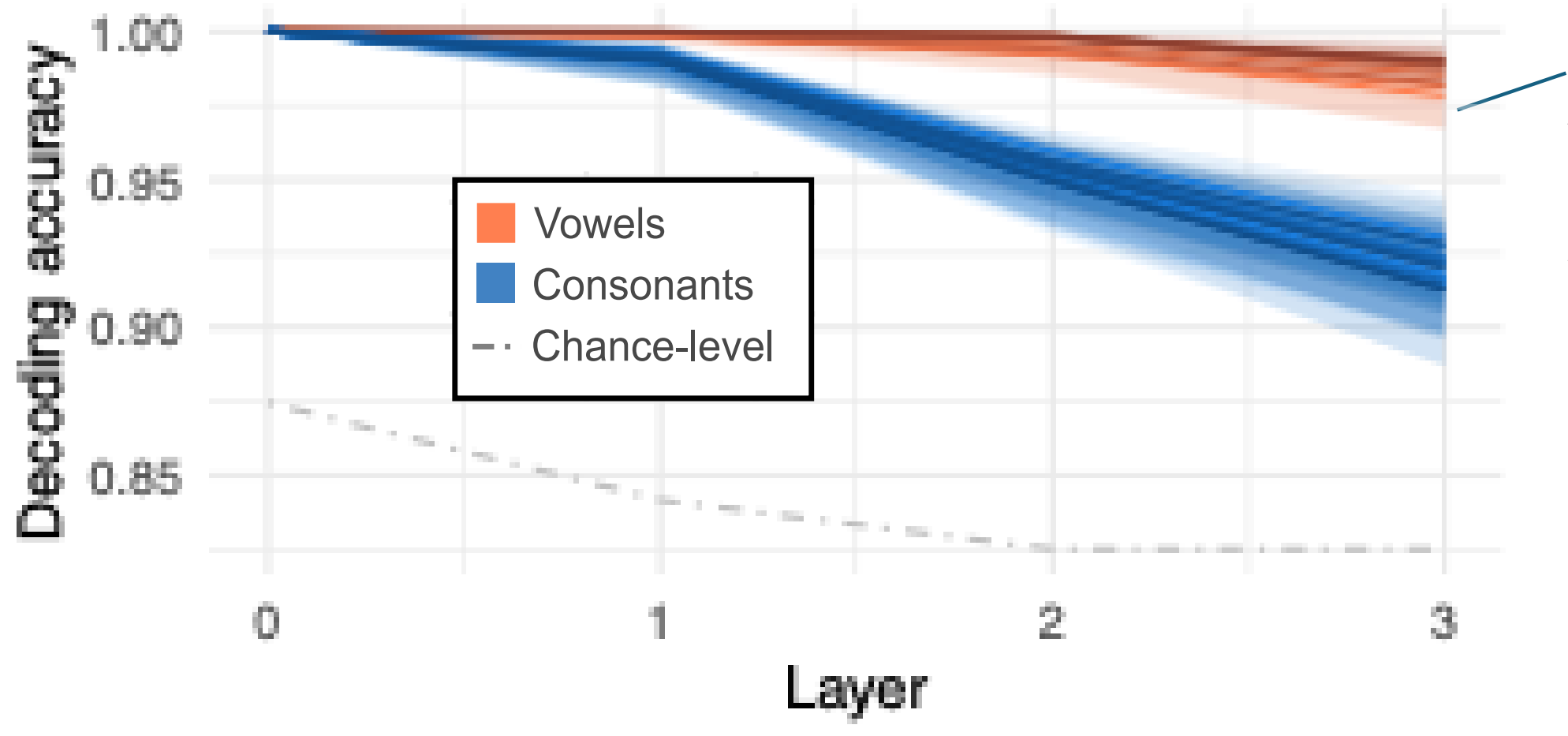
- **[l] vs. [r]**, e.g., *splölüm* vs. *sprölüm*
- **[p] vs. [t] vs. [k]**, e.g., *s**p**rolum* vs. *s**t**rolum* vs. *s**k**rolum*

✷ These Cs were chosen because they have the same distribution in language ⇒ mitigate effects of string context.

Results: highly decodable even in final layers ⇒ rejects strong hypothesis.

When a CNN successfully learns a vowel harmony pattern, what is the learned algorithm? Does it resemble tier-based analyses of harmony? While layers of such a CNN do not directly correspond to tiers, we find hallmarks of tier-based representations suggesting a "soft" representation of tiers.

Prioritization of vowels over consonants



Each line means: Accuracy of a linear classifier fit to classify whether a specific phoneme is present in a CVCVC string.

E.g., for [t]: *torel* vs. *pikem*

- All phonemes are decoded far above chance across layers.
- Though, there is clear prioritization to represent the identity of each vowel over consonants.

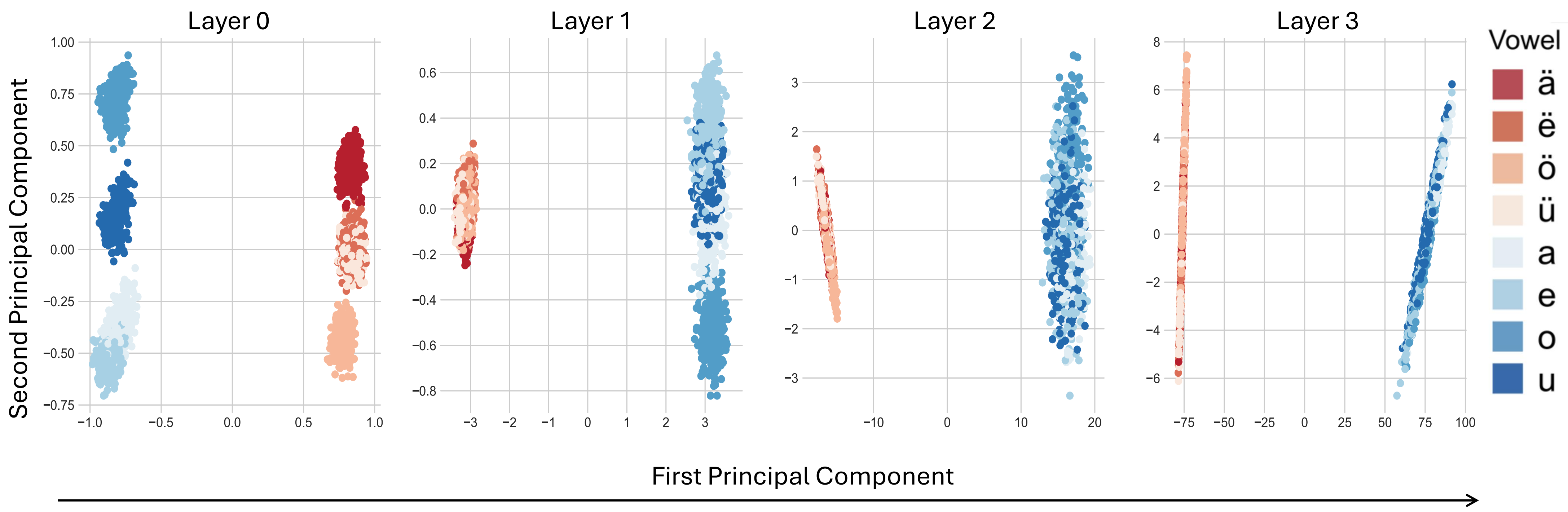
Feature abstraction over vowels

What strings have distinct/shared representations in the top principal components?

⇒ Informs us along which dimensions the model has learned to abstract over.

⇒ Specific instantiation of weak hypothesis: gradual (over layers) soft isolation of vowels.

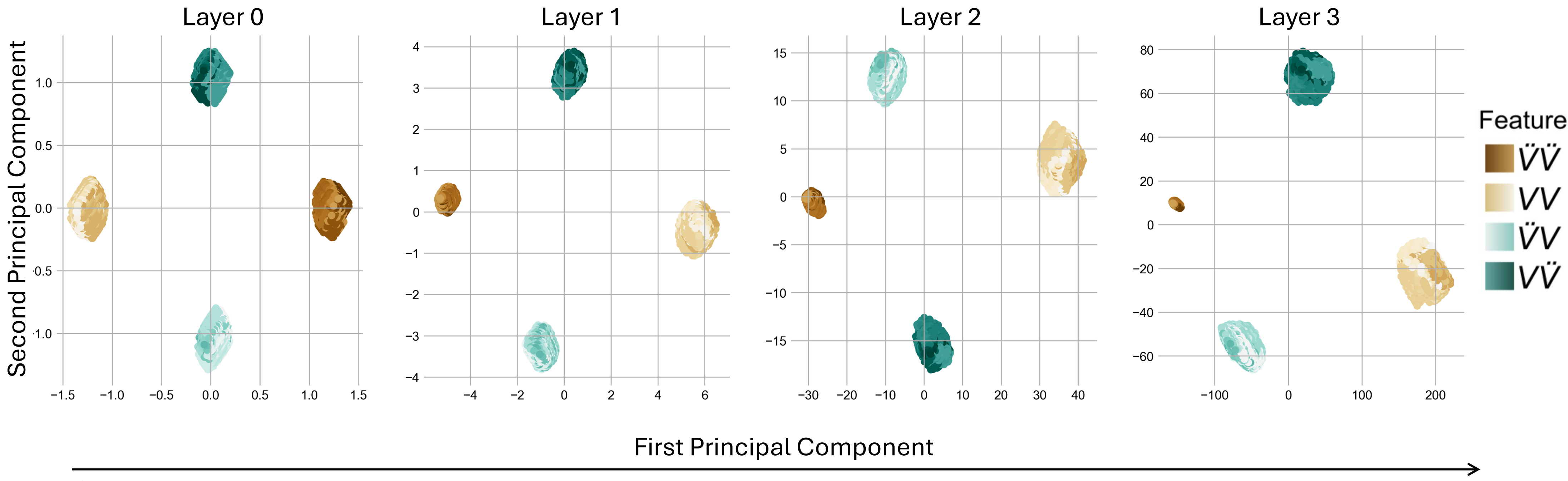
CVC string representations, coloured by identity of the vowel



Separation between [+umlaut] and [−umlaut] segments →

Abstraction within [±umlaut] group, identity of V progressively obscured →

CVCVC string representations, coloured by identity of V1V2 sequence



Separation between [±u][±u] groups, specific VV identity is obscured, CCC identity obscured, no merging into (un)grammatical groups

Discussion

We don't find evidence for a 1-to-1 relation between layer and tiers, but some important properties of tier representations have emerged in these CNNs.

Why no neutralization of Cs?

- Keep track of likely V positions. *p*(*c*|*V*) not uniform across c ∈ *C*.
- PCA results inform us that even if there are distinctions among Cs, they are represented in later PCs.
- Note: neutralization of Cs would be unexpected (and bad) as a learner of a phonological system, but in this toy example it was superfluous.

On the relation between model and humans

- Too early to draw connections to humans from this study.
- If similar results are obtained for CNNs learning other harmony patterns: the biases that CNNs introduce are the types of biases that a human might need to represent long-distance phonological dependencies.

Follow-up / addressable confounds

- Decodable representations ≠ used in critical computations ⇒ Intervention or encoding analysis.
- Effect of task and input ⇒ Similar analyses for phoneme models or models acquiring representations from raw speech
- Separation between grammaticality representations and tier-like representations.

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

[1] Goldsmith (1976). PhD thesis **[2]** Heinz, Rawal, & Tanner (2011). *ACL*.
[3] Odden (1994). *Language*. **[4]** Nevins (2010). *Locality in Vowel Harmony*.