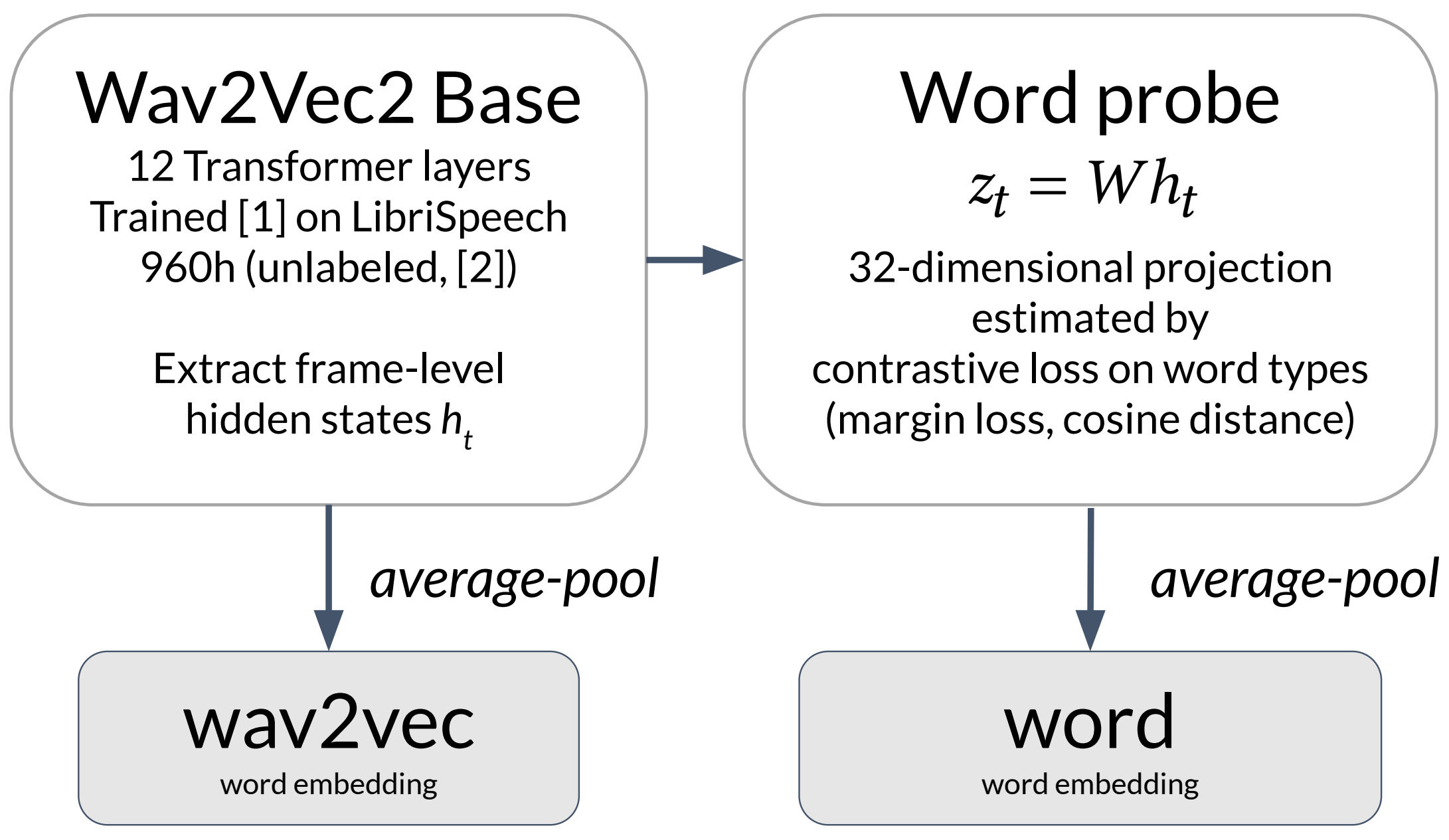
# Self-supervised speech models contain linear morpho-phonological structures

What linguistic representations are necessary to **recognize spoken words**? We study acoustic word embeddings of two models:



Wav2Vec2 Base
12 Transformer layers
Trained [1] on LibriSpeech 960h (unlabeled, [2])
Extract frame-level hidden states $h_t$

Word probe
$z_t = Wh_t$
32-dimensional projection estimated by contrastive loss on word types (margin loss, cosine distance)

*average-pool*

wav2vec
word embedding

*average-pool*

word
word embedding

## APPROACH

Study model encodings of word-final /z/, /s/, and /ɪz/, which are governed by overlapping morphological and phonological processes:

| | Allomorph | Base | Inflected |
|---|---|---|---|
| **Noun plural** | /z/ | daughter | daughters |
| | /s/ | lip | lips |
| | /ɪz/ | age | ages |
| **Verb 3rd-person singular** | /z/ | bring | brings |
| | /s/ | speak | speaks |
| | /ɪz/ | please | pleases |

Does a regular geometry link these word pairs in acoustic word embedding space? Is this geometry sensitive to **phonological** or **morphological** distinctions?

daughters : daughter :: speaks : speak

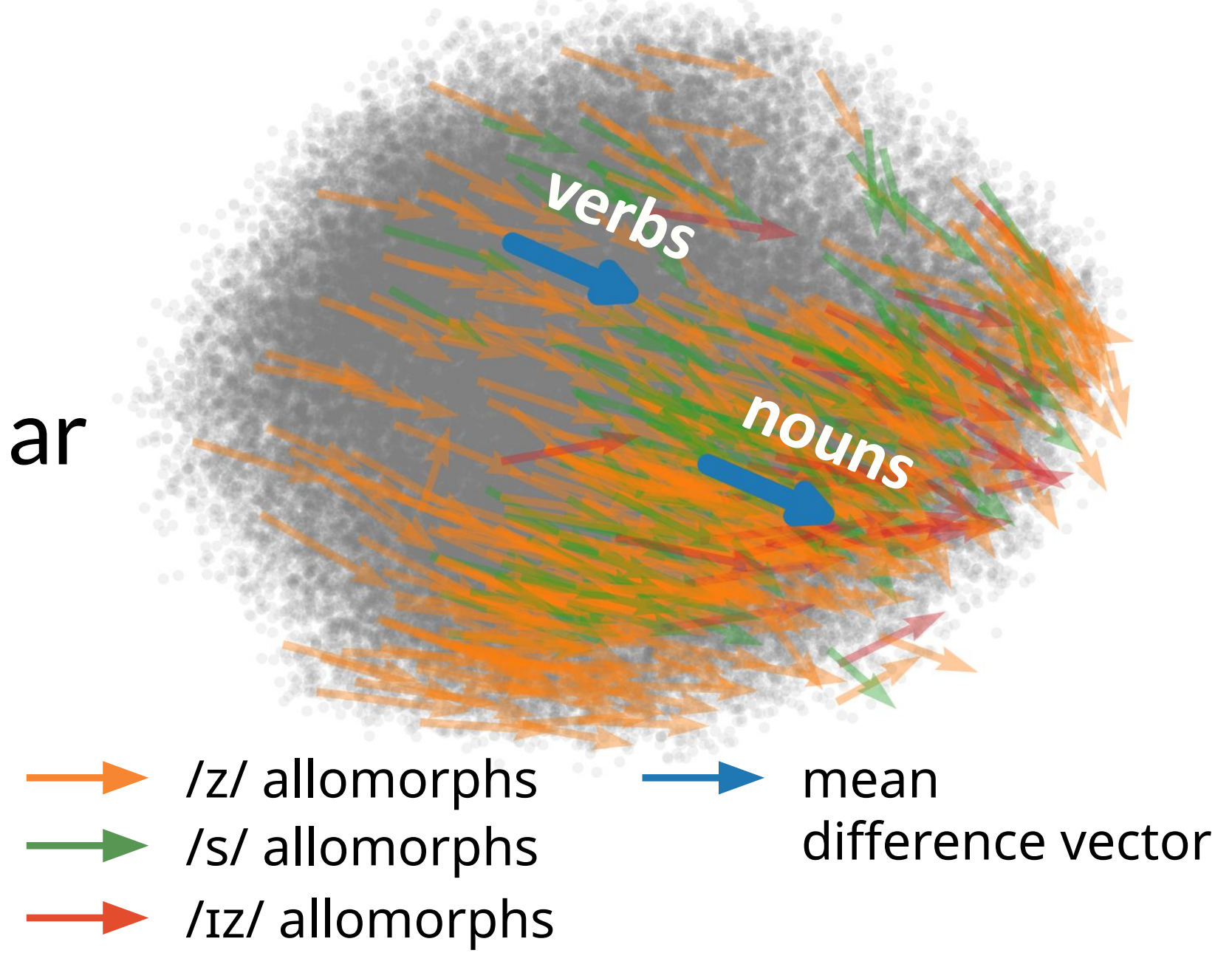
$$f(\text{daughters}) - f(\text{daughter}) \approx f(\text{speaks}) - f(\text{speak})$$

Predict analogy solutions from token-level eml $\hat{y} = f(\text{daughters}) - f(\text{daughter}) + f(\text{speak})$

Evaluate predictions by **rank** of the nearest valid

## RESULTS

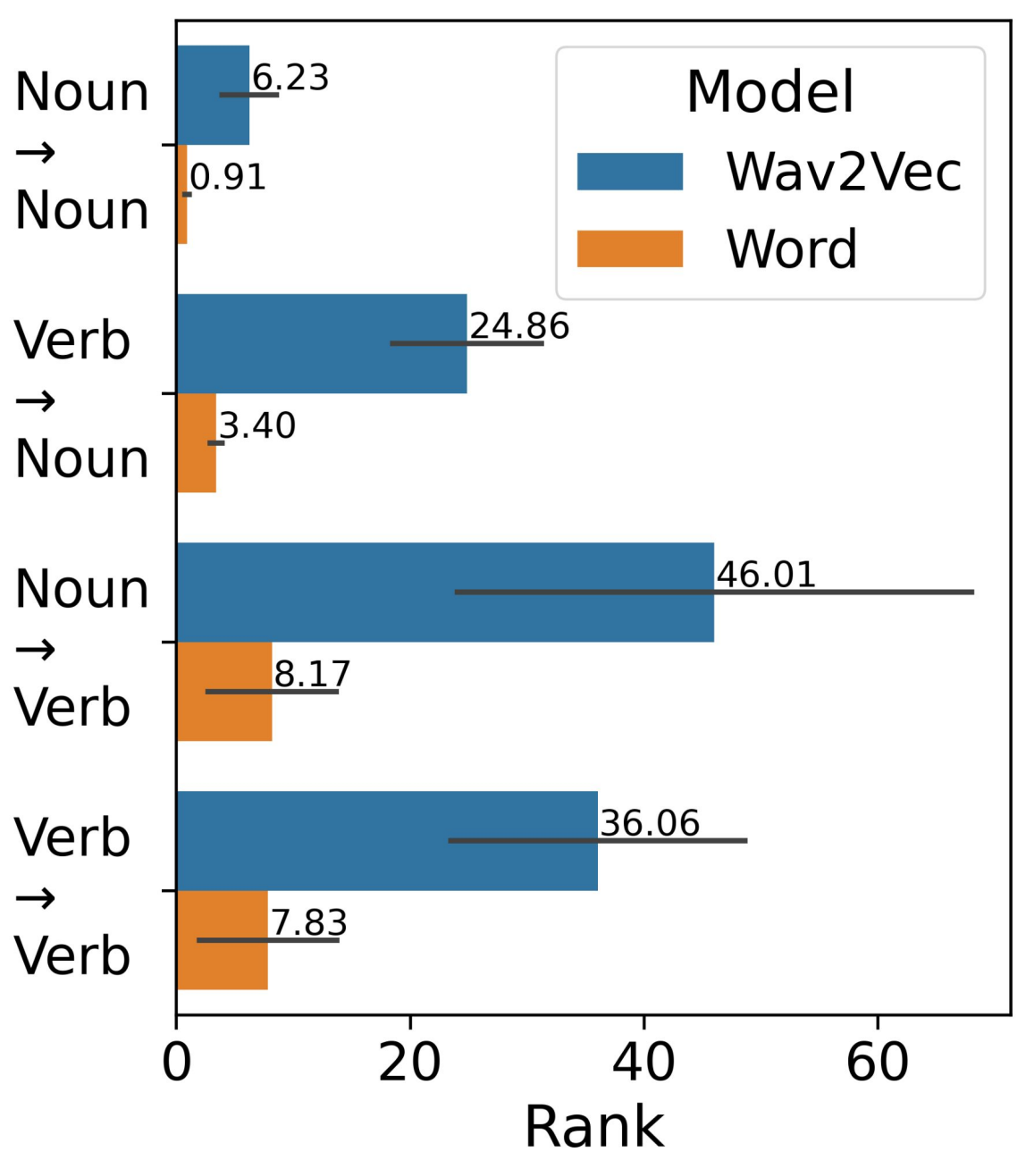
- Both embedding spaces show a regular direction linking base and inflected forms



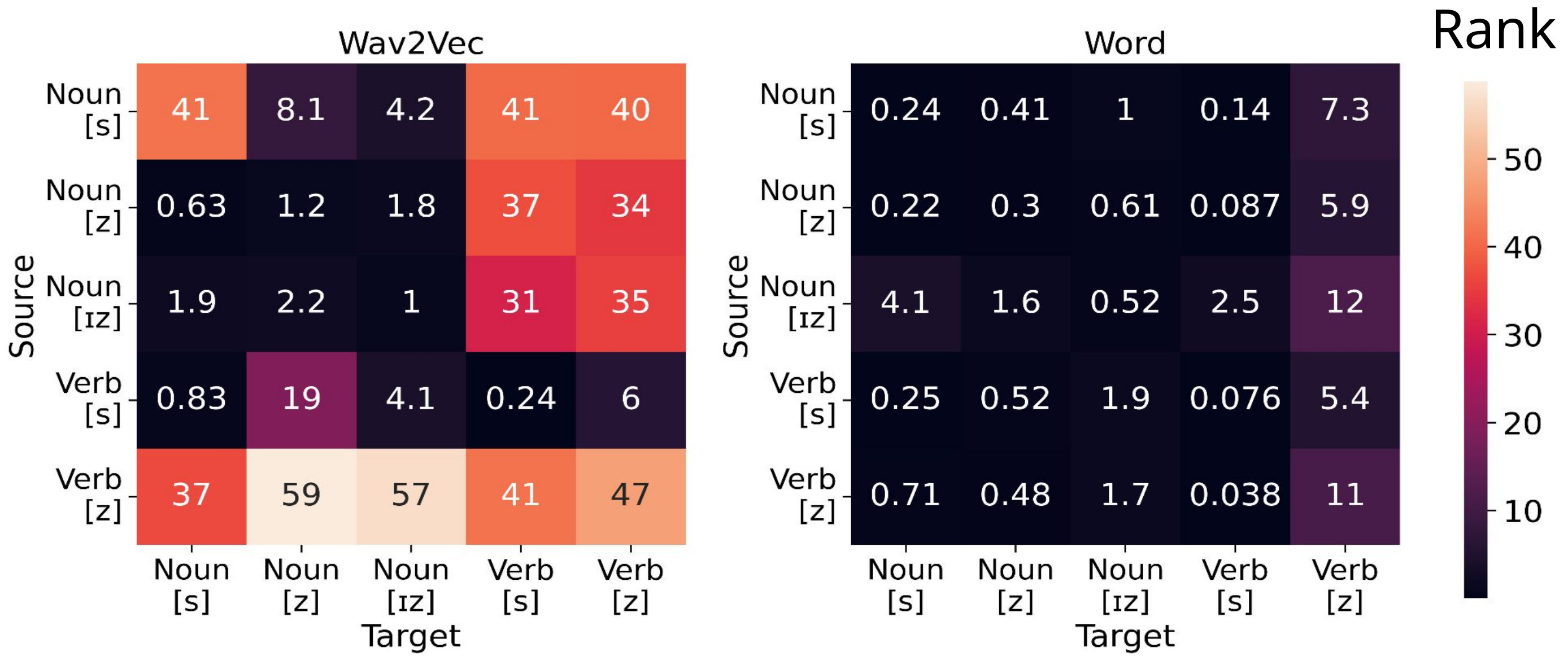
- Relations in the fine-tuned word embedding space generalize across both **morphological** and **phonological** distinctions; this is not true of the pretrained Wav2Vec embeddings

### Morphological generalization

```
war : wars :: lip : ________
Noun            Noun

speak : speaks :: lip : _____
Verb                Noun

war : wars :: exist : _______
Noun            Verb

speak : speaks :: exist : ___
Verb                  Verb
```

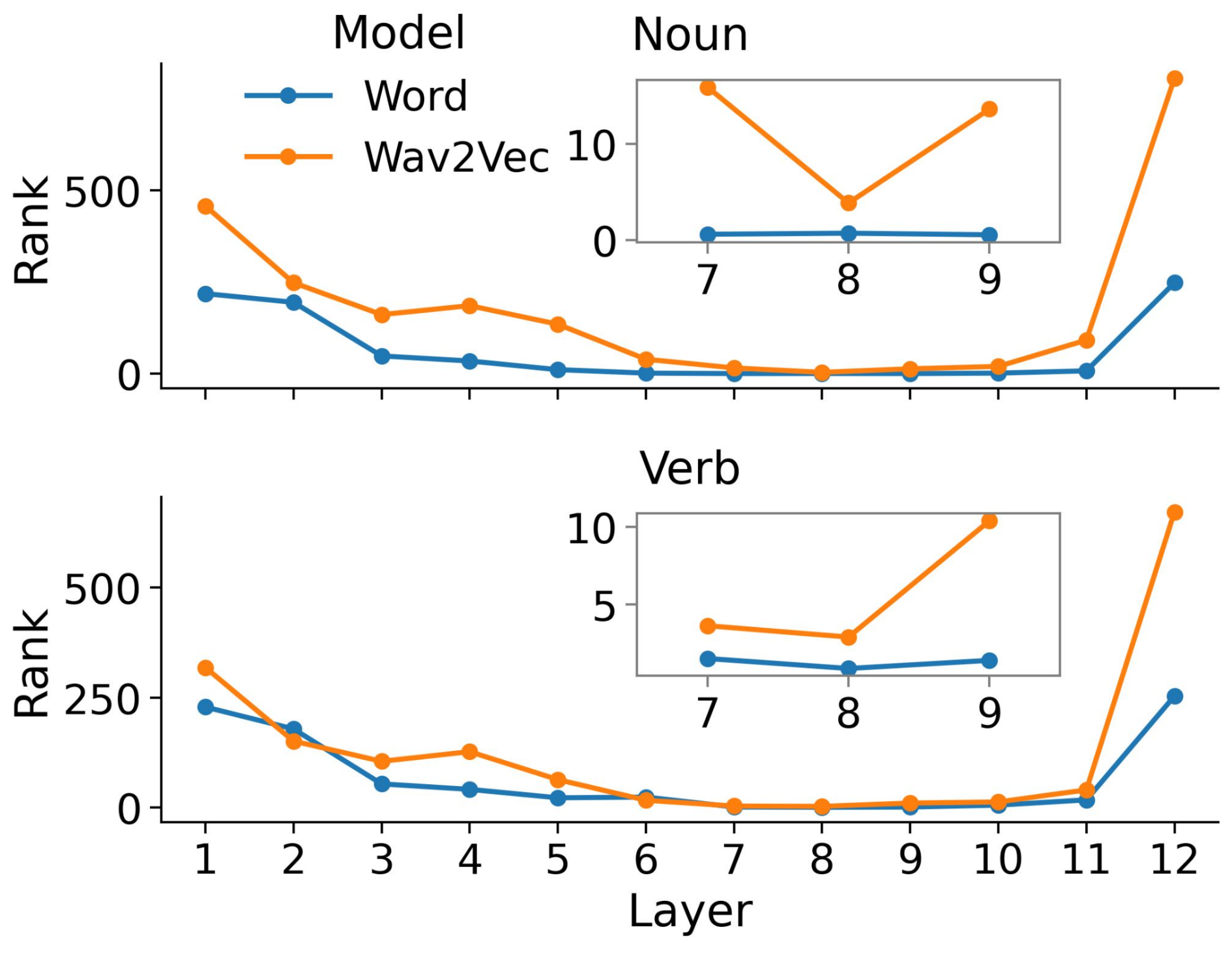


### Phonological generalization



**Per layer results**

Results above use the highest performing layer for both models (layer 8).



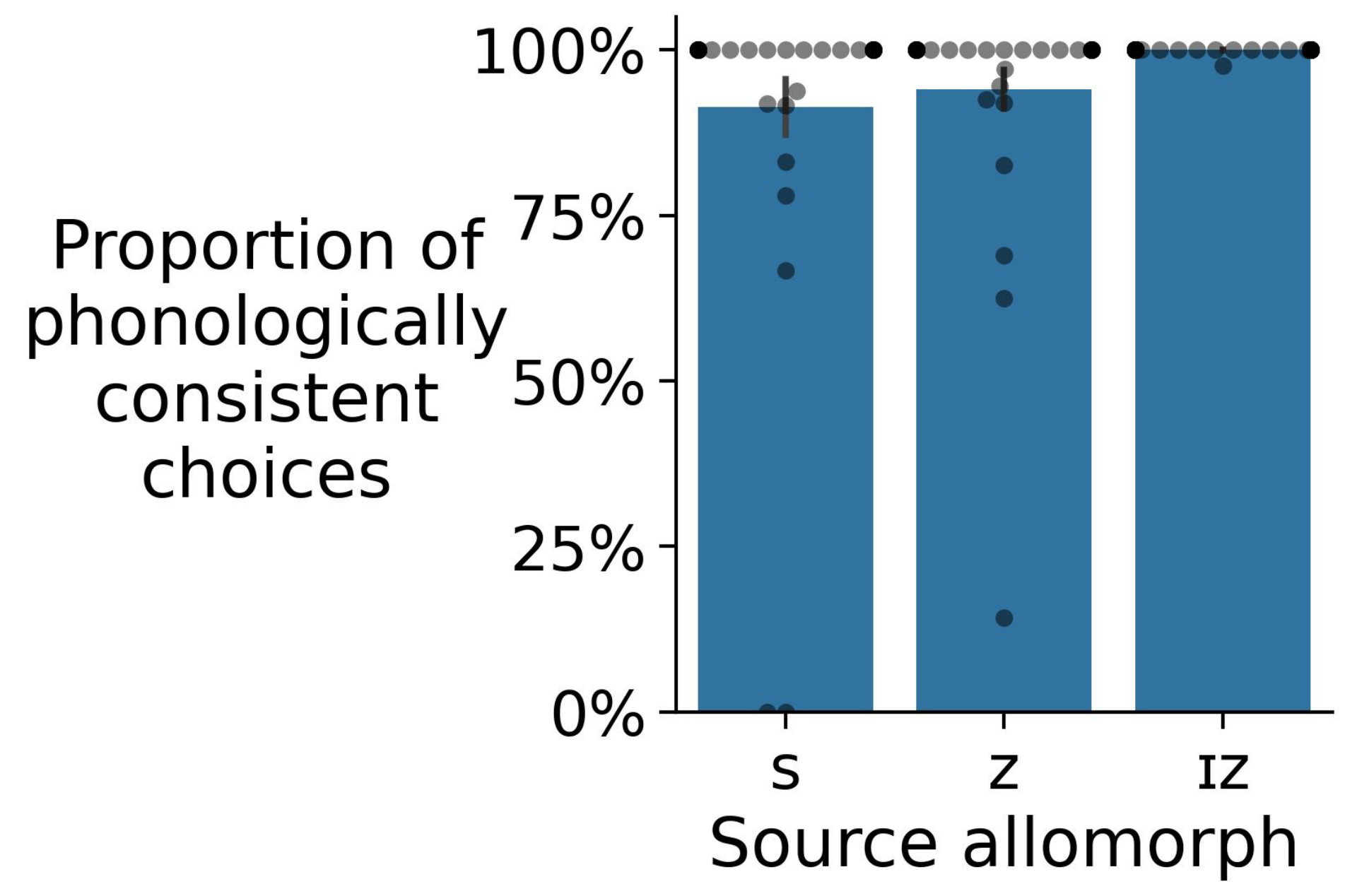
**English phonological conditions** on verb and noun inflections

/ɪz/ after sibilants
*dishes* [dɪʃɪz], *pleases* [plizɪz]
/z/ after voiced sounds
*dogs* [dɔgz], *brings* [bɹɪŋz]
/s/ after voiceless sounds
*cats* [kæts], *lips* [lɪps]

## FORCED-CHOICE ANALYSIS

This movement in embedding space effectively adds a sound which is **consistent with the phonological rules** exhibited in noun and verb inflections.

```
  [z]
own : owns :: bay : { bays  (consistent)   }
                    { base  (inconsistent) }
```



## TAKEAWAYS

- Acoustic word embeddings contain a linear subspace tracking an abstract morpho-phonological relationship between words
- This subspace is discoverable by optimizing for **word contrast**
- These results are not sensitive to embedding method (see paper)
- This suggests possible representations supporting human spoken word recognition

**Citations**

**[1]** Baevski et al. NeurIPS 2020.
**[2]** Panayotov et al. ICASSP 2015.

**Jon Gauthier, Canaan Breiss,**
Matthew Leonard, Edward Chang

University of California
San Francisco

USC University of Southern California

Poster, slides, and code available here