Emergent morpho-phonological representations in self-supervised speech models

Jon Gauthier[1] **Canaan Breiss**[2,3] **Matthew Leonard**[1] **Edward F. Chang**[1]
[1]Department of Neurological Surgery, University of California, San Francisco
[2]Department of Linguistics, University of Southern California
[3]Center for Computational Language Sciences, University of Southern California
jon@gauthiers.net cbreiss@usc.edu

Abstract

Self-supervised speech models can be trained to efficiently recognize spoken words in naturalistic, noisy environments. However, we do not understand the types of linguistic representations these models use to accomplish this task. To address this question, we study how S3M variants optimized for word recognition represent phonological and morphological phenomena in frequent English noun and verb inflections. We find that their representations exhibit a global linear geometry which can be used to link English nouns and verbs to their regular inflected forms.

This geometric structure does not directly track phonological or morphological units. Instead, it tracks the regular distributional relationships linking many word pairs in the English lexicon—often, but not always, due to morphological inflection. These findings point to candidate representational strategies that may support human spoken word recognition, challenging the presumed necessity of distinct linguistic representations of phonology and morphology.

1 Introduction

In everyday speech perception, humans transform a sequence of rapidly articulated words, often at a rate of several words per second, into coherent meanings. Traditional psycholinguistic models attempt to explain this behavior by positing distinct representations of perceived speech at several linguistic levels: listeners processing individual phonetic segments in a word, along with the phonological, morphological, and semantic properties of a word. This transduction from low- to high-level representation happens through a series of recurrent computations (McClelland and Elman, 1986; Gaskell and Marslen-Wilson, 1997).

While these models successfully explain a host of phenomena within the psycholinguistics of word recognition, they operate at a scale far below today's self-supervised speech models (S3Ms). Modern speech recognition algorithms that use S3Ms can make context-sensitive transcriptions of noisy acoustic input with an accuracy, efficiency, and scale not achievable with these psychologically motivated models. It is thus necessary to reconcile the theoretical commitments of earlier psycholinguistic models with the behavioral superiority of S3Ms.

Recent work has begun to study the representations of S3Ms trained on unlabeled speech data, asking to what extent they recover traditional psycholinguistic notions in phonetics, phonology, morphology, and syntax (Dunbar et al., 2022; Pasad et al., 2023b,a; Martin et al., 2023; Sanabria et al., 2023; Choi et al., 2024, 2025). As a psycholinguistic account, this work departs from prior approaches by searching for model capacities which emerge as a solution to a training objective, rather than seeking to build in hypothesized structures.

However, it is not clear how these capacities of S3Ms relate—if at all—to the function of word recognition. Because S3Ms are trained on broad objectives unrelated to any specific linguistic task, their internal states plausibly serve multiple functions beyond recognizing words. As such, we do not know which components of their learned representations are *necessary* for recognizing words, as opposed to predicting other aspects of the speech signal (from speaker identity and prosody to the identities of individual speech segments).

We present an analysis method and a series of controlled experiments on S3Ms, separating representations which are necessary for word recognition from those which emerge in service of more general objectives. We first design a probing method to operationalize the notion of optimal word recognition. This probe targets a subspace of an S3M which serves to distinguish spoken words. In a series of experiments studying the activations within this word-optimal subspace, we discover

Analysis code available at https://anonymous.4open.science/r/862E/.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28056–28075
November 4-9, 2025

/z/ → [z] after voiced non-sibilant sounds (*dogs* [dɔgz̲], *runs* [ɹʌnz̲])

/z/ → [s] after voiceless non-sibilant sounds (*cats* [kæts̲], *jumps* [dʒʌmps̲])

/z/ → [ɪz] after sibilant sounds (*dishes* [dɪʃɪz̲], *finishes* [fɪnɪʃɪz̲])

Box 1: Distributional constraints describing how the word-final /-z/ of English noun plurals and third-person verb inflections are realized in different surface forms [z], [s], and [ɪz].

a highly regular representation of speech sounds. We argue that this representation defies the cleanly separated levels of prior psycholinguistic models, cross-cutting morphological and simpler phonological distinctions, and instead tracks a higher-level pattern that unites multiple morphological inflections that all conform to a single phonological distribution rule. Our findings motivate a new candidate representational strategy that may serve spoken word recognition in human listeners.

2 Background

Like many languages, English exhibits regular sound patterns which reflect an interaction of phonological, morphological, and lexical constraints. For example, English words ending in [z], [s], or [ɪz] can reflect plural nouns (NNS; e.g. *dogs*, *cats*) and third-person singular verbs (VBZ; e.g. *runs*, *barks*). Both of these inflections introduce a morpheme with the underlying phonological content /-z/ at the right edge of their base, which is realized as one of three surface forms (*allomorphs*) depending on a simple set of distributional constraints, given in Box 1.

Of course, not all word-final instances of [z], [s], and [ɪz] are generated by these morpho-phonological processes: consider monomorphemic words such as *haze*, *fleece*, *six*, and *hearse*. Some of these words happen to end in sequences that are consistent with the constraints of Box 1, while some do not: *haze* and *six* end in a sound matching the voicing of the preceding sound (consistent); *fleece* and *hearse* end in sounds which do not match the voicing of the preceding sound (inconsistent).

A combination of morphological, phonological, and lexical processes produce these word-final sounds. We exploit these multiple levels of patterning in a series of experiments on a model of word recognition, asking how representations at these levels are negotiated by the model in service of its objective.

3 Methods

We design experiments targeting four (non-exclusive) hypotheses about the linguistic representations in self-supervised speech models. From model activations computed on the word-final sounds [z], [s], and [ɪz], our experiments evaluate:

Morphological sensitivity (§4.1): A morphologically sensitive representation of [z] would contrast instances of the plural [z] (e.g. in *daughters*) from instances of third-person singular [z] (in *enters*); likewise for [s] and [ɪz].

Phonological sensitivity (§4.2): A phonologically sensitive representation of [z] would contrast instances where it surfaces as [z] (e.g. *daughters*) from instances where it surfaces as [s] and [ɪz] (e.g. *lips* and *cheeses*); and likewise for [s] and [ɪz] in both noun plurals in verb inflections.

Lexical sensitivity (§4.3): A lexically sensitive representation of [z] would contrast instances where the [z] is part of a lexical form (e.g. *haze*) from instances where it is inflectional (e.g. *daughters*); and likewise for [s] and [ɪz].

Distributional sensitivity (§4.4): A representation sensitive to the distributional constraints given in Box 1 would contrast sounds [z], [s] and [ɪz] in contexts where they are consistent with these rules (e.g. *haze*) from contexts where they are inconsistent (e.g. *fleece*).

We address these questions using two types of models: a self-supervised speech model trained with a general contrastive learning objective, and a fine-tuned variant trained specifically for word recognition. By comparing answers to the above questions across these two models, we can assess which levels of linguistic representation are actually necessary for performing word recognition.

3.1 Base model

We begin our analyses with the Wav2Vec2 Base model (Baevski et al., 2020, herein Wav2Vec)[1],

[1] huggingface.co/facebook/wav2vec2-base

a self-supervised Transformer model of raw audio, which was trained on 960 hours of unlabeled data from the LibriSpeech corpus (Panayotov et al., 2015, publicly available under a CC-BY 4.0 license). This Transformer model takes an audio waveform as input and produces frame representations $x_\ell^{(t)}$, a sequence of 768-dimensional vectors each spanning 20 ms of audio beginning at time t, arranged in a series of model layers ℓ. While this variant of Wav2Vec was trained with entirely unlabeled audio, without phone- or word-level annotations, these models still accumulate detailed high-level representations of the linguistic input in individual frames (Pasad et al., 2023a).

3.2 Word probe model

What is the subspace of these Wav2Vec representations which optimally contrasts spoken words? We target this subspace by defining a linear projection on Wav2Vec's representation $x_\ell^{(t)}$ at frame t and layer ℓ onto a vector $z_\ell^{(t)}$:

$$z_\ell^{(t)} = W_z x_\ell^{(t)} \tag{1}$$

This probe is optimized with a contrastive learning objective. For each frame t within the span of a word j, we take all other frames t^+ spanned by other tokens of j as positive examples, and all other frames t^- spanned by tokens of distinct words as negative examples. We minimize a hinge loss, describing the separation in cosine distance between a frame and its positive and negative examples with margin parameter m:

$$\mathcal{L}(t) = \max\Big(0, m + \cos(z_\ell^{(t)}, z_\ell^{(t^+)}) - \cos(z_\ell^{(t)}, z_\ell^{(t^-)})\Big) \tag{2}$$

We train this probe on 100 hours of word-aligned audio from LibriSpeech from the split `train-clean-100`.[2] This model is trained to convergence separately for each layer ℓ of the Wav2Vec2 base model (12 layers in total). Further optimization details are included in Appendix B.5.

We take the activations of this model $z_\ell^{(t)}$ to be an optimal subspace of Wav2Vec for performing word-level contrast. By analyzing these activations, we can understand which aspects of the speech input are exploited for word recognition. By comparing these activations to the original Wav2Vec activations $x_\ell^{(t)}$, we can identify what information is present in the input but discarded by the model when performing word-level contrast.

[2]Alignments available in Lugosch (2019).

Inflection	Allomorph	Base	Inflected
Noun plural (NNS /z/)	[z]	*daughter*	*daughters*
	[s]	*lip*	*lips*
	[ɪz]	*cheese*	*cheeses*
Verb 3SG (VBZ /z/)	[z]	*give*	*gives*
	[s]	*exist*	*exists*
	[ɪz]	*please*	*pleases*

Table 1: Examples of unambiguous regular inflections.

3.3 Acoustic word embeddings

For all word tokens w_j in the word-annotated LibriSpeech corpus, we compute a fixed-length word representation by averaging across the N_j frame representations between the word's onset time o_j and offset (Sanabria et al., 2023; Pasad et al., 2023a).[3]

$$f_\ell(w_j) = \frac{1}{N_j} \sum_{t=0}^{N_j-1} z_\ell^{(o_j+t)} \tag{3}$$

These frame representations may correspond to the hidden states of Wav2Vec at layer ℓ, $x_\ell^{(t)}$, or the word probe $z_\ell^{(t)}$.

3.4 Experiments

Our experiments use the vector analogy method of Mikolov et al. (2013) to study the model's ability to generalize its representations of the relevant speech sounds [z], [s], and [ɪz] to lexically, phonologically, and morphologically contrasting strings.

Our analogy trials link two pairs of words, each of which are phonologically identical but for a final [z], [s], or [ɪz]. Table 1 gives examples of such pairs consistent with either noun pluralization or verb inflection. An example analogy is as follows:

shirt [ʃɚt] : *shirts* [ʃɚts] :: *cheese* [ʧiz]: _____

We implement this analogy with vector algebra. For random samples of word tokens a = *shirt*, b = *shirts*, c = *cheese*, we compute word

[3]While our experiments center around these word-level embeddings, this method is not critical to our results. Appendix B.3 shows that our results hold under a different embedding and analogy method, using representations pooled within individual phonemes rather than across entire words.

token representations $f_\ell(a_i), f_\ell(b_i), f_\ell(c_i)$, following Equation (3). We randomly draw token word embeddings a_i, b_i, c_i and calculate:

$$\hat{d}_i = b_i - a_i + c_i \quad (4)$$

For each trial, we compute the cosine distance between the predicted vector $\hat{d}_i$ and all word embeddings $f_\ell(w_j)$ computed across the LibriSpeech corpus. We average these predicted distances d_i across instances of the given analogy, and use these to compute a single nearest-neighbors ordering over all word tokens. This yields a single ranking over all word tokens for each analogy structure (a, b, c).

We evaluate with a rank metric: in the nearest-neighbors list for a given analogy, we find the position of the target word d. This rank value ranges from 0 to $334,008$ (the number of word tokens in the dataset). We define a random baseline performance for an analogy pair (a, b, c, d) by sampling a random counterfactual pair of words $(\overline{a}, \overline{b})$ and using the same method to compute the rank of d in the neighbors of predicted vectors $\overline{b}_i - \overline{a}_i + c_i$.

3.5 Experimental design

We select 753 frequent English nouns and 52 verbs which were present in the LibriSpeech dataset and are unambiguous with respect to part of speech: that is, their inflected form is clearly a noun or a verb.[4] While these unambiguous pairs share no actual morphological relationship (they instantiate an unambiguous plural noun or verb inflection), they do share a phonological relationship, all exhibiting inflected forms with an underlying word-final -/z/.

4 Results

We first evaluate the overall capacity of each layer of our word probe, along with Wav2Vec, to solve analogy tasks across these English noun and verb inflections. Figure 1 plots results at each layer; model predictions exceed random chance (rank $= 94,306$) for all inflections at all layers.

These layer-wise results mirror the typical pattern of abstract feature encoding in S3Ms, with maximal performance in intermediate layers and a substantial decrease in performance in final layers

[4]This is not true of most English nouns: for example, the form *stacks* can be either the plural of the noun *stack* or the third-person singular of the verb *to stack*. We also exclude words which are technically unambiguous, but which have base or inflected forms that are homophonous with a monomorpheme: for example, *patients–patience*; *knows–nose*. Additional details on word selection are in Appendix A.

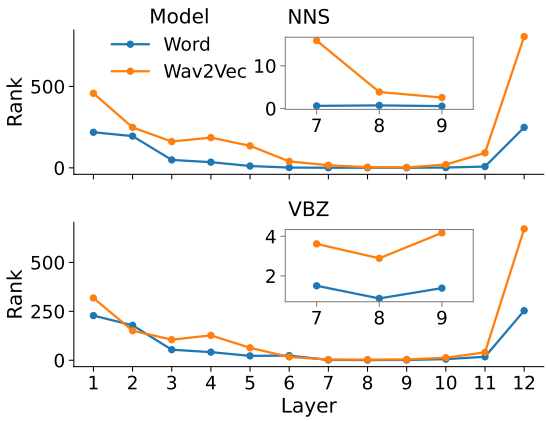

Figure 1: Per-layer analogy rank performance (lower is better) for regular noun and verb inflections. Random chance is $94,306$. Insets show zoom on layers 7–9.

closest to the model's prediction head. The performance of Wav2Vec peaks in the 8th and 9th layer, matching the peak location of phonetic encoding found in prior studies (Pasad et al., 2023b).

Both Wav2Vec and the word probe thus exhibit a global linear geometry that links inflected nouns and verbs to their base forms. This geometry is visible in the principal component space of all acoustic word embeddings, visualized in Figure 2.

What kind of linguistic information is captured by these difference vectors? By design, the analogy trials we construct frequently link source and target pairs which are mismatched in morphological category (noun and verb inflections) and phonological detail (the particular allomorphs involved). The following experiments (summarized in Table 2) ask whether the particular morphological and phonological relationship between the source and target pairs affect the success of this analogy task. If analogy succeeds despite morphological and phonological mismatches, this indicates that the difference vectors are not dependent on distinctions at these levels. We present experiments on the highest-performing layer in Wav2Vec and the word probe (layer 8). We focus on high-level mean rank results in the main text, with more detailed quantitative analyses provided in Section 4.5 and Appendix B.1.

4.1 Evaluating morphological sensitivity

We first evaluate the model's ability to generalize within and across morphological categories, from plural noun inflections to both 1) other plural noun inflections and 2) third-person verb inflections, and vice-versa. If the representation supporting overall

	Method	Example	Answer
§4.1	Test analogy across **morphological** categories (NNS→VBZ, VBZ→NNS)	*daughter* : *daughters* :: *own* : *owns* /z/ — NNS /z/ — VBZ	Invariant
§4.2	Test analogy across **allomorphs** ([z], [s], [ɪz])	*daughter* : *daughters* :: *lip* : *lips* [z] [s]	Invariant
§4.3	Test analogy to/from false inflections	*daughter* : *daughters* :: *beside* : *besides*	Invariant
§4.4	Forced-choice analogy evaluation for **phonological consistency**	*daughter* : *daughters* :: *bay* : { *bays* (consistent) / *base* (inconsistent) }	Prefers sounds consistent with Box 1

Table 2: Summary of experimental designs and results on the word probe.

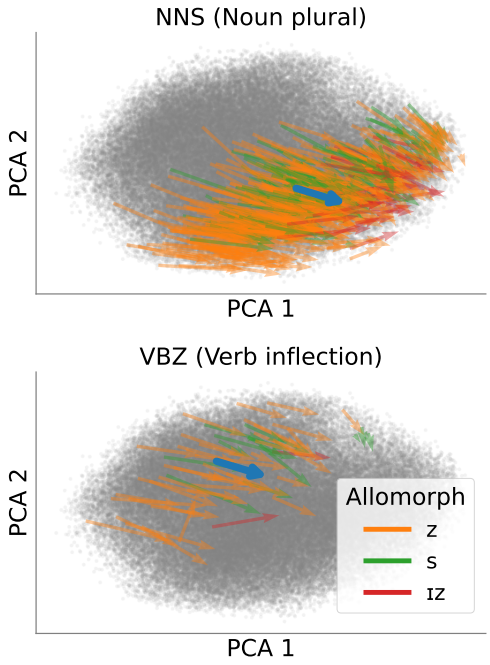

Figure 2: Difference vectors from base forms a to inflected forms b, computed on layer 8 of the word probe and projected into the first two principal components of embedding space for all words in our study. Bold blue line shows mean direction vector; gray dots show a random sample of word embeddings.

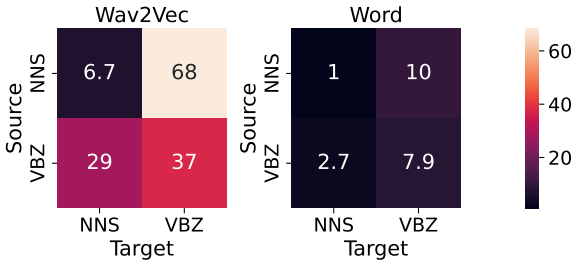

Figure 3: Mean rank values (lower is better) from analogy within and between noun/verb inflections on the Wav2Vec baseline and word probe.

analogy performance is primarily morphological, this latter evaluation should fail, since by design there is no shared morphological relationship between our unambiguous noun pairs and verb pairs. If the representation does not depend on morphological facts, we should see success in both cases.

Figure 3 shows the results of this analogy evaluation. In each heatmap, the diagonal elements show the average rank values resulting from analogies within-inflection (e.g. *shirt*/*shirts* to *cheese*/*cheeses*), while the off-diagonal elements show the results of performing analogies between inflections (e.g. *shirt*/*shirts* to *enter*/*enters*).

We first examine the results of the word probe (right heatmap). By comparing the NNS→NNS cell (top left) and VBZ→VBZ cell (bottom right), we can see a main effect of morphology: while both categories perform far above chance, nouns are better predicted than verbs ($t = -4.01, p < 10^{-4}$).[5] The off-diagonal cells test analogy from noun plural inflections to verb inflections and vice-versa. This evaluation also performs well above chance, though we see a similar superiority for predicting noun inflections over verb inflections.

These results show a highly restricted role of morphology within the word probe's computations. This contrasts with the results of the same evaluation applied to Wav2Vec's internal states, shown in the left panel of Figure 3. Wav2Vec performs substantially worse overall ($t = 4.43, p < 10^{-5}$), and also shows a much larger sensitivity to morphological contrasts: while the word probe shows an average rank difference of mismatching morphology of 7.1, the same metric in the Wav2Vec results is 34.7. The word probe thus supports computations which

[5]Error analysis suggests that this main effect stems from differences in the size of noun and verb morphological paradigms. English nouns have a very small morphological paradigm primarily involving pluralization, while verbs have a larger paradigm (3SG, past, participle, gerund). A common error in 3SG verb prediction is selecting the wrong inflection within this paradigm, such as predicting *pleasing* instead of *pleases* from *please*.

Base	Inflected	Base	Inflected
backward	backwards	beside	besides
the ([ðiː])	these	though	those

Table 3: False friends of the most frequent allomorph, [z].

are relatively invariant to morphological contrasts.

4.2 Evaluating phonological sensitivity

We next addressed our question about sensitivity to phonological contrasts within and between each inflectional category. We did this by splitting the previous unambiguous noun–verb experiment based on the particular allomorph ([z], [s], or [ɪz]) involved in both the base and target pair.

Figure 4 shows the results of this evaluation applied to our probe model. Each cell value indicates the accuracy of the model in generalizing from some particular allomorph base to a different allomorph target pair. The Wav2Vec model (left panel) shows strong sensitivity to phonological and morphological distinctions: while analogy performance is almost as good as that of the word probe model for particular cases (e.g. VBZ [s] → NNS [s], *exist* : *exists* :: *lip* : *lips*), the majority of mappings show degraded performance (e.g. VBZ [s] → NNS [z], *exist* : *exists* :: *daughter* : *daughters*).

In contrast, the word probe model (right panel) is relatively stable across both phonological and morphological differences between the source and target of an analogy. We still see moderate effects of morphological identity (in particular, the row-wise effect of drawing NNS [ɪz] as a source, and the column-wise effect of generalizing to VBZ [z] as a target), but the scale of this degradation is far smaller than that of the Wav2Vec evaluation.

4.3 Evaluating lexical sensitivity

Our results show that Wav2Vec's encoding of word-final [z], [s], and [ɪz] is sensitive to both the morphological and phonological conditions distributing these sounds. These details are discarded, however, by a model optimized for word recognition. We next ask if this pattern extends beyond the morphological phenomena of noun plurals and verb inflections, to distinct *lexical* sources of word-final [z], [s], and [ɪz]. We derive "false friend" word pairs, which on their surface appear just like English noun and verb inflections: the "inflected" form is the concatenation of a "base" form with one of the sounds [z], [s], or [ɪz], obeying the rules given in Box 1. Table 3 shows examples of such false friends attested in English.[6]

Figure 5 evaluates how difference vectors computed from true noun and verb inflections perform in predicting in false friend (FF) items, and vice versa, in both Wav2Vec and the word probe. The top-left cell of each panel indicates the analogy performance within valid noun and verb inflections; note that these values correspond to the diagonal values of Figure 3. The off-diagonals indicate the ability of the model to compute analogies linking valid inflections with false friends (for example, the top right cell of left panels, NNS→NNS-FF, tests analogies such as *shirt* : *shirts* :: *though* : *those*).

The Wav2Vec model (top panels) shows a strong sensitivity to real versus false-friend inflections, as is visible on the diagonals of the heatmaps: noun analogy prediction degrades from a rank of 6.7 to 76, and verbs from 37 to 66. In contrast, we see a much smaller change in the word probe (lower panels), with a change in rank outcome in nouns from 1 to 8.4 and in verbs from 7.9 to 8.7.

This suggests that the word probe is relatively insensitive to the contrast between lexical and morphological sources of word-final [z], [s], and [ɪz]. It must instead rely on the phonological relationship between these words' base and inflected forms.

4.4 Evaluating distributional sensitivity

But what kind of phonological relationship is encoded in these difference vectors linking spoken words? The allomorphy results of Figure 4 show that the word probe model is apparently invariant to distinctions in English speech sounds which are contrastive: sounds which *must* be distinctly represented in order to distinguish English words.

Concretely, consider the minimal pair of *bays* and *base*, both of which can be derived by the addition of a single sound to the word *bay*.[7] While *bays* is consistent with the rules given in Box 1—it contains a word-final [z] following a voiced sound—*base* is not consistent, placing a word-final [s] after a voiced sound. We have seen that the word probe is largely insensitive to the particular word-final sounds in these words. However, it is still possible

[6] We exclude all true nouns and verbs as possible false friends, since any analogy evaluation relating real nouns and verbs would confound the phonological "false friend" relationship with other morphological relationships.

[7] The attentive reader will also note that these words may contrast in vowel length. This confound is addressed in Appendix B.6.

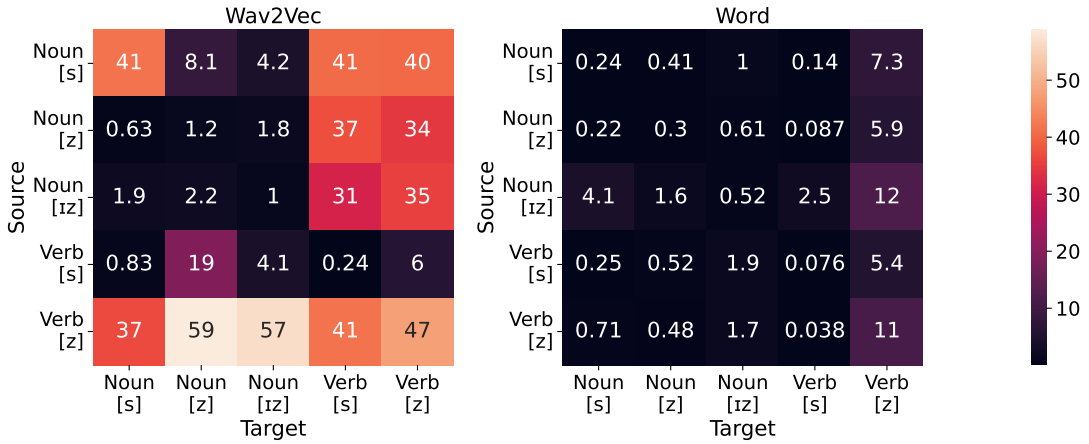

Figure 4: Analogy within and between allomorphs of noun and verb inflections. Heatmaps show an average rank metric (0 is best; random guessing is $94,306$). The word probe in the right panel exhibits strongly reduced sensitivity (i.e., improved performance) to allomorphic contrasts.

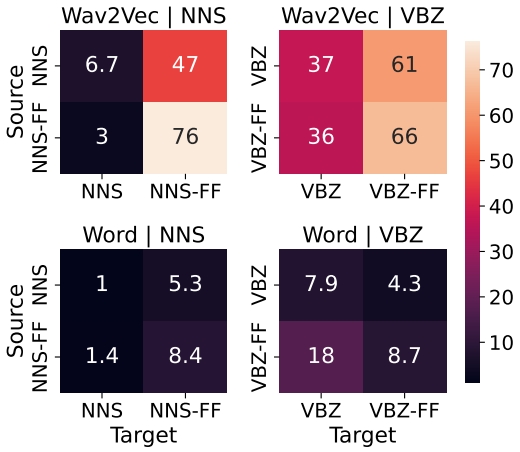

Figure 5: Analogy to false-friend (FF) inflection pairs. Word probe (bottom panels) shows reduced sensitivity to whether or not sounds are participating in true morphological inflections.

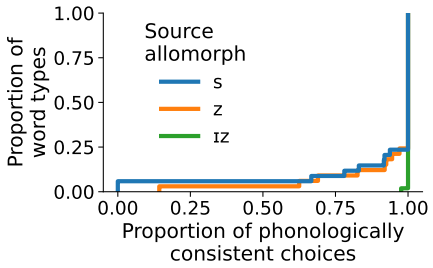

Figure 6: Cumulative distribution of preferences for the item obeying distributional constraints (Box 1) in forced-choice evaluations on 35 target pairs. The majority of forced-choice evaluations predict the consistent item 100% of the time.

that the word probe is sensitive to the *distributional constraints* of these sounds: whether or not they are consistent with the kind of rules shown in Box 1.

In a final experiment, we select 35 word triples (Table 10) exhibiting the following property: the "base" is a substring of both other words; the consistent word is the concatenation of the base with a [s], [z], or [ɪz], following the distributional constraints of Box 1; the inconsistent word is the concatenation of the base with a sound not following these constraints.[8] We perform a forced-choice analogy evaluation, mapping difference vectors from real noun and verb inflections onto the base word, e.g.:

$$lip : lips :: bay : \left\{ \begin{array}{l} bays \text{ (consistent)} \\ base \text{ (inconsistent)} \end{array} \right\}$$

We measure the probability that the model predicts a vector $\hat{d}_i$ (eq. (4)) whose nearest neighbor is a token of the phonologically consistent form *bays* rather than the inconsistent form *base*. We find that the word probe's difference vectors overwhelmingly point to the consistent option of each forced-choice item, with a majority of items showing an absolute preference for consistency (Figures 6 and 12). This suggests that the word probe representations are not completely invariant to contrasts between these sounds. Instead, they capture the abstract distributional constraints that give rise to [z], [s], and [ɪz] in regular noun and verb inflections.

4.5 Regression evaluation

Previous sections claimed that the word probe showed a substantial reduction in its sensitivity to

[8] Some of the phonologically consistent items are also morphologically related to the base (*bay–bays* is a plural inflection), while others have no relationship at all (*knew–news*).

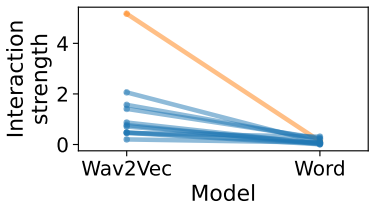

Figure 7: Interaction strength (a regression-based measure of sensitivity to morphological and allomorphic distinctions) in Wav2Vec and the word probe.

morphological (sections 4.1 and 4.3) and phonological (section 4.2) contrasts, relative to the Wav2Vec baseline model. We now quantitatively test these claims with a linear regression, predicting the rank outcome of individual analogy trials as a function of the allomorphs and morphological categories involved, along with nuisance predictors:

$$\begin{aligned}\texttt{rank} \sim\ & \texttt{allomorph\_from} \times \texttt{inflection\_from} \times \\ & \texttt{allomorph\_to} \times \texttt{inflection\_to} + \\ & \texttt{from\_frequency} + \texttt{to\_frequency} \end{aligned} \quad (5)$$

where `inflection` features are categorically coded in {`NNS`, `VBZ`}, `allomorph` features are categorically coded in {`s`, `z`, `Iz`}, and `frequency` features are log-transformed frequencies from Worldlex (Gimenes and New, 2016). This model is fit independently on the results of Wav2Vec and the word probe, at the same model layer studied in previous sections.

The interaction coefficients of these regressions capture how sensitive the models are to matches or mismatches between the morphology and allomorphy of source and target pairs. Figure 7 tracks the absolute value of each interaction coefficient explaining the performance of the Wav2Vec model and word probe. For example, the strongest interaction coefficient in the Wav2Vec fit (left point of the orange line in Figure 7) captures the large difference between noun-to-verb and verb-to-verb analogies in the right column of Figure 3. This coefficient is greatly reduced in the word model (right end of orange line), along with that of all other interaction effects.

This reduction in interaction effects is consistent with our earlier findings: optimizing a model for word recognition produces representations with substantially reduced sensitivity to both allomorphic and morphological variation. Appendix B.1.1 presents the full results of the regression evaluation.

5 Discussion

This paper used speech representations computed from self-supervised speech models (S3M) to ask: what kinds of representations are necessary for recognizing words? We studied how an S3M variant optimized for word recognition, the word probe, negotiates phonological, morphological, and lexical levels of representation to serve this objective.

Our experiments show that the model represents an abstract phonological regularity in English: a highly frequent *distributional regularity* (Box 1) governs word-final [s], [z], and [ɪz]. This knowledge is not strongly conditioned by morphology (sections 4.1 and 4.3) or phonological relationships between allomorphs (section 4.2), and instead tracks a morpho-phonological process governing the sounds which form English noun plurals and present tense verb inflections (section 4.4). This generalization emerges early in the model (as early as layer 1) and peaks in intermediate layers (fig. 1). In the word probe, this knowledge is encoded as a global linear translation, prominent in the principal components of the embedding space (fig. 2).

Our results add to prior probing studies of self-supervised speech models. Pasad et al. (2023b) found that intermediate layers 8 and 9 of Wav2Vec most prominently encoded phonetic information. Our word probe results show that these layers also contain a linear subspace capable of more abstract phonological reasoning; on the surface, however, these representations are indeed sensitive to phonetic contrasts. Choi et al. (2025) demonstrate that Wav2Vec's internal states retain sub-phonemic and allomorphic information; we confirm this allomorphic sensitivity in Section 4.2, and show that optimizing for an objective of word recognition effectively erases much of that allophonic information (Figure 3).

Our results also complement existing psycholinguistic models of spoken word recognition (McClelland and Elman, 1986; Norris, 1994; Gaskell and Marslen-Wilson, 1997, inter alia), which largely were designed to explain small-scale phenomena of single-word recognition. Due to this data scale, these past models required strong inductive biases about the kinds of intermediate linguistic computations necessary for word recognition. In contrast, by combining large amounts of naturalistic data, low-bias neural network models, and controlled experiments, we can uncover novel hypotheses about the computations underlying spoken

word recognition.

This kind of level-crossing speech representation is broadly consistent with some alternative distributed models of spoken word recognition in psycholinguistics, which emphasize how apparent rule-based morphological relationships can be modulated by linguistic knowledge at other levels of representation (Gonnerman et al., 2007). It is also compatible with a view in the theory of Distributed Morphology, known as the *separation hypothesis*, which similarly argues that representations of groups of phonologically distributed sounds (such as the set [z], [s], and [ɪz]) may be disconnected from the their original morphological sources (Embick et al., 2022).

Acknowledgments

We thank Connor Mayer, David Embick, members of the Chang Lab at UCSF and the Stanford GLySN lab, and the audience at the Bay Area Language Processing Interest Group for feedback on this work. Generative AI tools (ChatGPT 4o) were used to improve the wording and clarity of the manuscript.

Limitations

This paper focuses on a small (albeit highly frequent) set of morphological and phonological phenomena in English. Future work should ask whether this linear geometry applies to other English phenomena with dual patterning of morphology and phonology — for example, word-final *-er*, which is generated by both comparative (*bigger*) and agentive (*buyer*) inflections. We might also use multilingual or monolingual non-English models to verify that the patterns revealed in our experiments are truly a generalization resulting from exposure to English phonology specifically.

Our analyses are limited to a very strict assumed geometric relationship between word forms (linear translation). Our findings may be sensitive to this geometric assumption, yielding both false positives (on the sensitivity of Wav2Vec to morphology and phonology) and false negatives (on the lack of sensitivity of the word probe model).

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-supervised speech representations are more phonetic than semantic. In *Proc. Interspeech 2024*, pages 4578–4582.

Kwanghee Choi, Eunjung Yeo, Kalvin Chang, Shinji Watanabe, and David R Mortensen. 2025. Leveraging allophony in self-supervised speech models for atypical pronunciation assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2613–2628, Albuquerque, New Mexico. Association for Computational Linguistics.

Ewan Dunbar, Nicolas Hamilakis, and Emmanuel Dupoux. 2022. Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1211–1226.

David Embick, Ava Creemers, and Amy J Goodwin Davies. 2022. Morphology and the mental lexicon: Three questions about decomposition.

M Gareth Gaskell and William D Marslen-Wilson. 1997. Integrating form and meaning: A distributed model of speech perception. *Language and cognitive Processes*, 12(5-6):613–656.

Manuel Gimenes and Boris New. 2016. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior research methods*, 48:963–972.

Laura M Gonnerman, Mark S Seidenberg, and Elaine S Andersen. 2007. Graded semantic and phonological similarity effects in priming: evidence for a distributed connectionist approach to morphology. *Journal of experimental psychology: General*, 136(2):323.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Loren Lugosch. 2019. Librispeech alignments.

Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Levy. 2023. Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration. *arXiv preprint arXiv:2306.06232*.

James L McClelland and Jeffrey L Elman. 1986. The trace model of speech perception. *Cognitive psychology*, 18(1):1–86.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Dennis Norris. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. 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.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2023a. What do self-supervised speech models know about words? *arXiv preprint arXiv:2307.00162*.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023b. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Gordon E. Peterson and Ilse Lehiste. 1960. Duration of syllable nuclei in english. *Journal of the Acoustical Society of America*, 32:693–703.

Ramon Sanabria, Hao Tang, and Sharon Goldwater. 2023. Analyzing acoustic word embeddings from pre-trained self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Samuel A. Zimmerman and Stanley M. Sapon. 1958. Note on vowel duration seen cross-linguistically. *Journal of the Acoustical Society of America*, 30:152–153.

A Stimulus selection

We select 753 frequent English nouns and 52 verbs which were present in the LibriSpeech dataset and were *unambiguous* in their part of speech: that is, their inflected form can either be a noun or a verb, but not both.

Specifically, for each word type in the dataset, we calculated the distribution of its part-of-speech labels in the corpus, tagged with spaCy's `en_core_web_trf` part-of-speech tagger. We retained only those noun types which had no attested verb instances, and vice versa. We then manually excluded residual ambiguous items from the resulting list. Figure 8 shows the frequency distribution of the retained nouns and verbs and compares it with the marginal noun/verb frequency distribution in the corpus. Our stimulus filtering procedure does not select for long-tail rare words on which the models may have poorly calibrated representations — it seems to do just the opposite, selecting a relatively high-frequency subsample of word types.

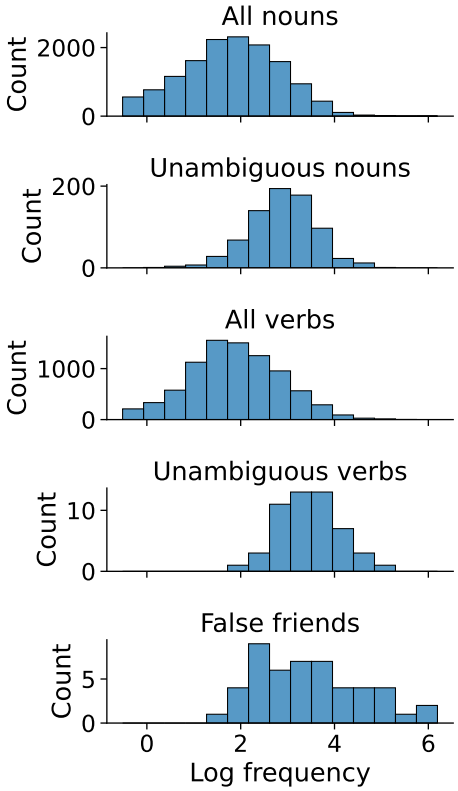

Figure 8: Frequency distributions of word stimuli used in our experiments. Each facet shows word log-10-frequency distributions on a different subset of words in the LibriSpeech corpus (`librispeech-train-clean-100` split). First two facets compare all nouns with the unambiguous nouns used in our experiments; the next two facets do the same for verbs. Final facet plots the frequency distributions of false friend items, used in Section 4.3.

Tables 4 and 6 give the complete list of unambiguous nouns and verbs used in our experiments.

B Supplementary results

B.1 Extended results of main analyses

Figures 9 to 12 offer more statistically detailed versions of the main plots Figures 3 to 6, respectively.

Median rank values are given for the equivalent mean rank visualizations in the main text: Figure 16 (analogous to Figure 3) and Figure 13 (analogous to Figure 4).

B.1.1 Regression results

Table 7 gives the full regression models estimated independently on the rank outcomes from the Wav2Vec and Word models. The interaction strength values of Figure 7 are generated by comparing the difference of absolute values of all rows

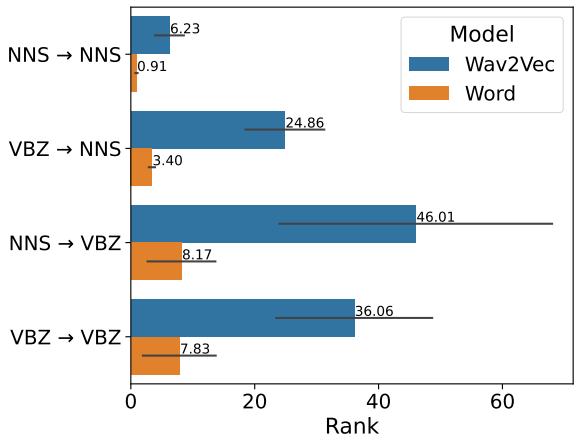

Figure 9: Rank mean and standard error estimates for the main morphology analysis of Section 4.1 and fig. 3. Mean values differ slightly from fig. 3 because this figure estimates performance pooled within single target word pairs in order to derive a meaningful uncertainty measure, whereas fig. 3 pools all trials together regardless of target word pair.

of this table which are interactions (i.e., terms including "×").

B.2 Results on held-out LibriSpeech data

The results reported in the main text are evaluated on the `librispeech-train-clean-100` data split of the LibriSpeech corpus, which was seen both during the pretraining of the Wav2Vec model and the word probe. To validate these findings, we also evaluated on the held-out split `librispeech-test-clean` set, which was not used during training or hyperparameter selection of either Wav2Vec or the word probe.

However, this test set is relatively small ($\sim$ 5 hours), leading to substantial data sparsity in our more granular analyses, such as the allomorphy evaluation of Section 4.2. For example, when we restrict our analysis to word types appearing in the `test-clean` split at least 5 times, there is exactly one regular unambiguous English noun plural usable in our analysis: *hearts*.

For this reason, the main text reports results on the much larger `train-clean-100` set, where we can evaluate these trends of interest. This section shows that results on the held-out `test-clean` set qualitatively—and, where measurable, quantitatively—replicate those reported in the main text. Figure 17 provides an analogous plot to Figure 3, Figure 14 to Figure 4, and Figure 19 to Figure 6. Some of the cells of Figure 14 are null due to data sparsity.

The quantitative trends given in the main text hold in the held-out dataset:

- Wav2Vec performs worse than the word probe on the analogy text ($t \approx 2.00, p \approx 0.0481$).
- Nouns are better predicted than verbs in analogies on the word probe representations ($t \approx -2.89, p \approx 0.00584$)

B.3 Results without word-level pooling

It is possible that the evaluation as given in the main text may disadvantage Wav2Vec relative to the word probe. Our word probe constrains all frames across the span of a word to converge to a single type-level representation, potentially erasing dynamic temporal information within Wav2Vec's frame representations (Section 3.2). This kind of code might be less sensitive to the average-pooling embedding method of Section 3.3 compared to the Wav2Vec model. It is possible, then, that our analogy evaluations are unfairly biased toward the word probe.

To address this, we introduce a *phoneme-level pooling* version of the experiments in the main text. In this setting, rather than submitting word embeddings (eq. (3)) to vector arithmetic (eq. (4)), we operate on individual *phoneme embeddings* from critical points within words.

For each regularly inflected word (e.g. *shirts*), we define a *constancy point* as the final phoneme it shares with its base form (i.e., the /t/ of *shirts*). Let $f_c(w_i)$ be the mean-pooled representation of the audio frames spanning the constancy point, and $f_f(w_i)$ be the mean-pooled representation of the final phoneme in the word (here /s/). Our difference vectors now compute the relationship $f_f(w_i) - f_c(w_i)$: the movement of the model representation from a point prior to the inflection to a point after the inflection.

To solve an analogy *shirt* : *shirts* :: *cheese* : _____, we perform an updated version of Equation (4) for some token i of *shirt* and some token j of *cheese*:

$$\hat{d} = f_f(\text{shirts}_i) - f_c(\text{shirts}_i) + f_c(\text{cheese}_j) \quad (6)$$

We qualitatively and quantitatively replicate the outcomes shown in the main text under this evaluation. Analogs of the main result figures are given in Figure 18 (analogous to Figure 3), Figure 15 (analogous to Figure 4), and Figure 20 (analogous to Figure 6). This confirms that the word probe

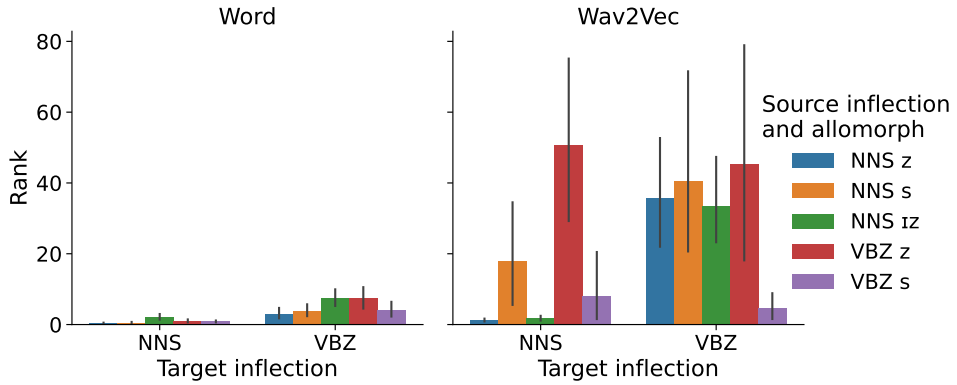

Figure 10: Rank mean and standard error estimates decomposed by both the morpheme and particular allomorph expressed in the source word pair.

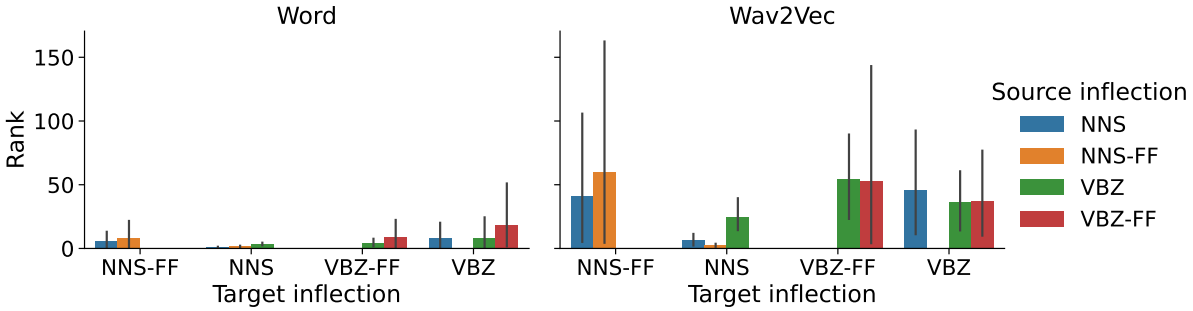

Figure 11: Rank mean and standard error estimates for the false friend analysis of Section 4.3 and fig. 5.

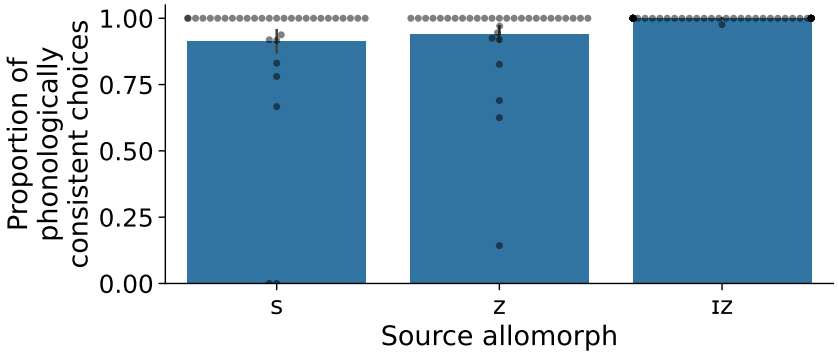

Figure 12: Proportions of phonologically consistent choices in the forced-choice experiment of Section 4.4. Each point corresponds to a forced-choice triple; results are grouped by the particular sound present in the source word pair.

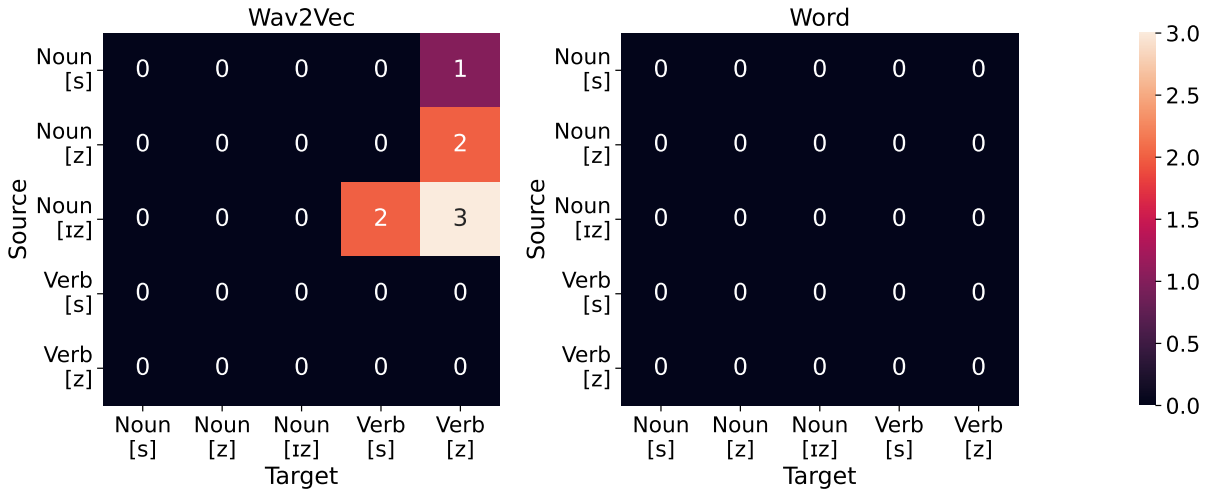

Figure 13: Median rank results for analogy within and between allomorphs of noun and verb inflections. Compare with mean values in Figure 4.

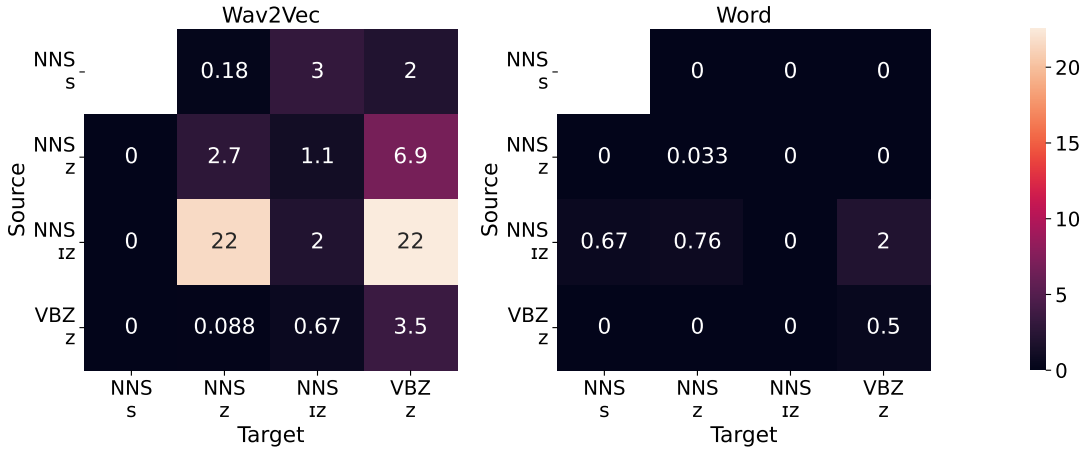

Figure 14: Analogy within and between allomorphs of noun and verb inflections, estimated on the data split `librispeech-test-clean` (see Appendix B.2). Heatmaps show an average rank metric (0 is best; random guessing is 94, 306). The word probe in the right panel exhibits strongly reduced sensitivity (i.e., improved performance) to allomorphic contrasts. Analogous to Figure 4.

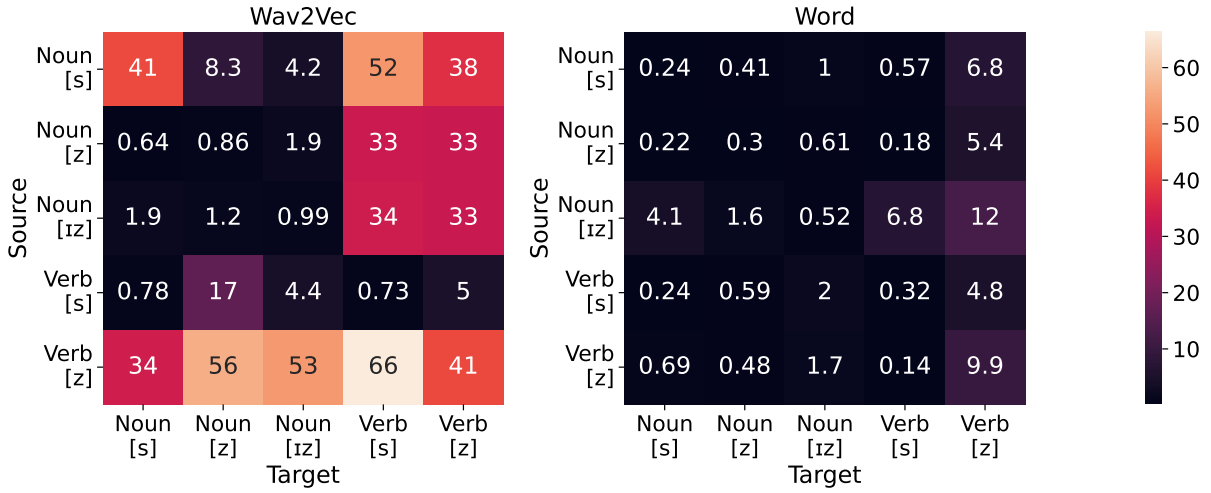

Figure 15: Phoneme-pooled analogy results within and between allomorphs of noun and verb inflections, estimated without word-level pooling (appendix B.3). Analogous to Figure 4.

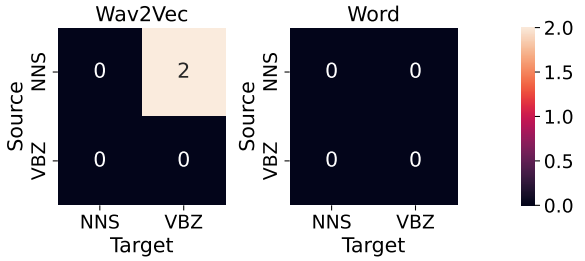

Figure 16: Median rank results for analogy within and between noun/verb inflections on the Wav2Vec baseline and word probe. Compare with mean values in Figure 3.

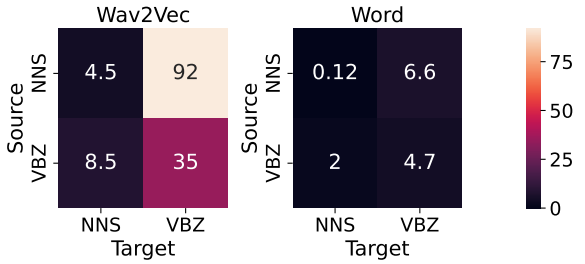

Figure 17: Mean rank values (lower is better) from analogy within and between noun/verb inflections on the Wav2Vec baseline and word probe, estimated on the data split `librispeech-test-clean` (see Appendix B.2). Analogous to Figure 3.

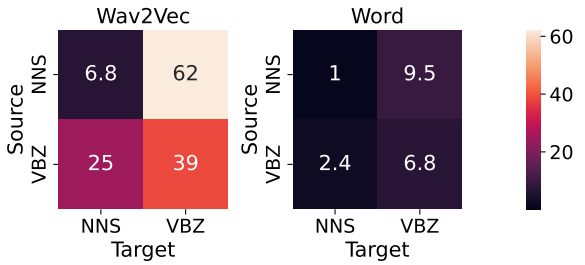

Figure 18: Phoneme-pooled analogy results within and between noun/verb inflection on the Wav2Vec baseline and word probe, estimated without word level pooling (appendix B.3). Analogous to Figure 3.

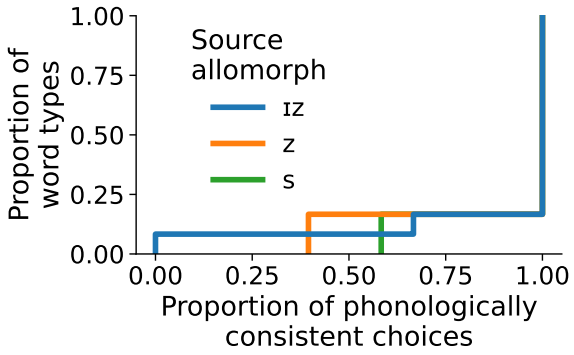

Figure 19: Cumulative distribution of preferences for the forced-choice item obeying distributional constraints, estimated on the data split `librispeech-test-clean` (see Appendix B.2). Analogous to Figure 6.

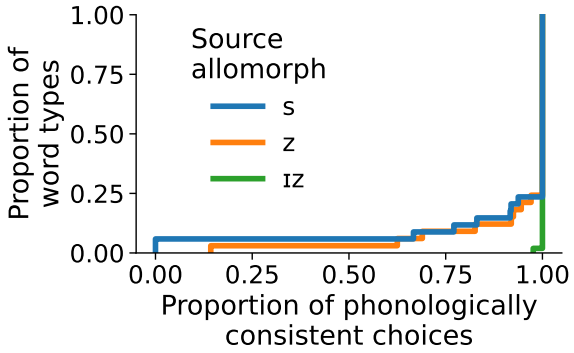

Figure 20: Cumulative distribution of preferences for the forced-choice item obeying distributional constraints, estimated without word-level pooling (Appendix B.3). Analogous to Figure 6.

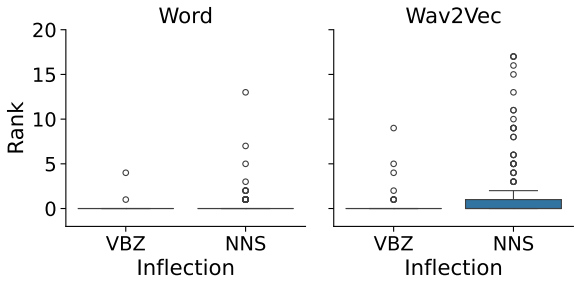

Figure 21: Rank outcomes of the same-word evaluation of Appendix B.4. Our analogy method can be applied to map between distinct tokens of single word types; this establishes a performance bound for the results of our main experiments in Section 3.4.

meaningfully alters the model's dynamic response to individual sounds, rather than simply generating a response which is more amenable to word-level average-pooling.

B.4 Same-word evaluation

Our analogy method asks whether a predicted word embedding $\hat{d}$ (Equation (4)) is close to any token of a desired word. For example, on the analogy *shirt* : *shirts* :: *cheese* : _____, we compute the minimum distance between the predicted $\hat{d}$ and any token of the desired word *cheeses*. This method can effectively capture the relationship between base and inflected forms of nouns and verbs in our main evaluations.

In a supplementary evaluation we asked whether this same logic applies across tokens of a single word type. If this were true, it would demonstrate that word types are **coherent** in embedding space, and establish a soft lower bound (maximal performance) for our transfer evaluations in the main text. We draw embeddings of arbitrary bases and their inflected forms x_{self} (e.g. *shirt*) and y_{self} (*shirts*), and split these into two disjoint "source" and "target" sets. We then ask whether difference vectors computed on embeddings from the source set can be transferred to embeddings from the target set:

$$\hat{d}_{\text{self}} = x_{\text{self}}^{\text{source}} - a_{\text{self}}^{\text{source}} + c_{\text{self}}^{\text{target}} \qquad (7)$$

We compare the predicted $\hat{d}_{\text{self}}$ to the unseen target embeddings $d_{\text{self}}^{\text{target}}$ and compute the same mean rank metric as described in Section 3.4. Mean rank outcomes are given in Table 8 for the Wav2Vec baseline and the word probe; Figure 21 compares outcomes of the same-word evaluation within nouns and verbs for the two models.

By comparing these values to the main results in Figure 3, we see that the analogy results across words in some cases reach the upper bound established by the same-word evaluation in both models. For example, the word probe performs best in noun–noun analogies with a mean rank of 1; slightly exceeding the model's same-word result of 1.9.

B.5 Model optimization

The word probe model computes a 32-dimensional representation of each frame through a linear transform of Wav2Vec's 768-dimensional activation (Equation (1)), with $32 \times 768 = 24,576$ total learnable parameters. We estimate individual probes for each layer of the Wav2Vec base model with AdamW (Loshchilov and Hutter, 2017), minimizing the hinge loss given in Equation (2) and early-stopping using a held-out validation dataset. Model hyperparameters (the margin parameter m, learning rate, and weight decay) are selected using a separate validation set in order to maximize the mean average precision of a classifier mapping from probe frames (Equation (1)) to word categories, following the word-annotated LibriSpeech corpus. Optimal values for the model analyzed in the main results of this paper (mapping from layer 8 of Wav2Vec) are given in Table 9.

A complete hyperparameter search for a single layer requires approximately 12 hours on a single NVIDIA TITAN V GPU.

B.6 Forced-choice experiment

Table 10 gives the full set of forced-choice triples used in Section 4.4.

Below we address several possible confounds that arise from this experimental design.

B.6.1 Possible confound: Bias for /z/

The experimental items only contain "consistent" words ending in [z]. It is unfortunately impossible to design a fully balanced experiment under this structure due to the limitations of English phonotactics. Any hypothetical "consistent" word ending in [s] would have by construction a preceding voiceless sound (e.g. *lips* [lɪps]). However, the addition of a voiced consonant here ([lɪpz]) is prohibited by more general constraints on English consonant cluster voicing.

At first glance, then, our results are also confounded with the simpler claim that these difference vectors function to add a [z] sound onto any

base form. However, we have shown that the same vectors exhibit roughly similar performance in predicting inflected forms with [s] and [ɪz] as with [z] in Section 4.2 and Figure 4 (compare results across columns).

B.6.2 Possible confound: Vowel length

English vowels are typically lengthened before voiced codas (Zimmerman and Sapon, 1958; Peterson and Lehiste, 1960). This raises a second possible confound: the difference vectors might be capturing changes in vowel length due to voicing, rather than (or in addition to) the presence of the final sibilant. However, this explanation doesn't fit with the results in Section 4.2 and fig. 4. Consider the first column of that figure, which tests whether noun plurals ending in [s] can be predicted using difference vectors from various sources. The second cell (row 2, column 1) uses vectors from words ending in [z] to predict inflections of words ending in [s]. If vowel length were driving the effect, we'd expect worse performance here—since words ending in [z] would exhibit vowel lengthening but words ending in [s] would not. Yet performance is similar to the case where both source and target words end in [s] (row 1, column 1), suggesting vowel length mismatch isn't a key factor here.

The regression results in Table 7 support this same idea: the estimated effect of sound match (`allomorph_from=S × allomorph_to=S`) in the word probe model is only an average 0.08 improvement in rank score.

Table 4: Unambiguous nouns (753) used in our experiments. Continued in Table 5.

ability	accident	accomplishment	achievement	acquaintance	acre	action	activity	actor
adder	advantage	adventure	advertiser	affair	affection	age	agent	agony
allusion	ambition	ancient	angel	angle	animal	ankle	apartment	ape
apple	application	apprehension	argument	army	arrangement	arrow	artery	article
artist	ash	assistant	association	attention	attitude	aunt	author	authority
baby	bag	baker	ball	band	bank	banker	banner	barn
barrel	barricade	barrier	basket	beast	beauty	bed	bee	beech
beggar	being	bell	bench	bird	biscuit	blade	blanket	boat
body	bond	bone	book	bottle	bough	boundary	boy	branch
breast	brother	brow	buccaneer	bud	buffalo	bull	bullet	bundle
burden	button	cabin	cage	cake	candle	cannon	canoe	canyon
captain	captive	car	card	carriage	cart	case	castle	cat
cave	cedar	cell	cent	century	ceremony	chair	chamber	champion
channel	chapter	characteristic	cheek	cheese	chicken	chief	chimney	church
circumstance	citizen	city	clerk	cliff	cloak	clock	cloud	club
coal	coat	coin	colony	column	combination	commander	commissioner	common
community	companion	company	competitor	complaint	composition	comrade	conception	conclusion
condition	connection	consideration	contemporary	contribution	conversation	cord	cordial	corn
corpse	cottage	counsel	country	couple	course	cousin	crane	creature
crime	criminal	crystal	cup	current	curtain	custom	customer	danger
daughter	day	death	debt	deck	degree	demonstration	depth	description
desk	detail	device	devil	diamond	difficulty	dinner	direction	disappointment
disaster	discovery	disease	doctor	doctrine	dog	dollar	door	doorway
dozen	drawer	duty	eagle	ear	eel	effort	egg	elbow
elder	element	emotion	enchantment	enemy	energy	engagement	engine	enterprise
error	estate	evening	event	evil	example	exception	exclamation	excursion
exertion	expectation	expense	experiment	expert	explanation	expression	extremity	fact
factor	factory	failure	fairy	family	farmer	father	feather	feature
fellow	female	field	finger	fir	fist	flag	fleet	flight
flood	floor	flower	folk	forest	fort	fortune	foundation	fountain
fowl	fragment	friend	frog	fruit	fund	fur	gale	game
garden	garment	gate	general	generation	gesture	ghost	giant	gift
girl	glacier	glass	glimpse	god	good	government	gown	grain
grape	grave	grove	guest	guinea	gun	habit	hair	hall
hardship	hat	heart	heaven	hedge	heel	height	hen	hero
hill	hip	historian	history	hole	holiday	home	horn	horror
horse	host	hotel	hour	house	humor	hundred	hunter	husband
hut	idea	ideal	illusion	image	imagination	impression	improvement	impulse
inch	incident	income	inconvenience	indication	individual	injury	insect	instinct
institution	instruction	instrument	intention	interruption	interval	investigation	invitation	island
jackal	jaw	jewel	joy	junior	justice	key	kind	king
kingdom	knee	knight	laborer	lad	lady	lake	lamp	lane
language	lantern	law	lawyer	leader	league	leg	lesson	letter
liberty	lily	limb	lion	lip	list	lord	loss	lot
lover	luxury	machine	magistrate	maiden	maker	manner	map	maple
marriage	martian	mast	material	maxim	meadow	meal	medicine	melody
member	memory	merchant	message	messenger	meter	method	mile	mill
million	mink	minute	miracle	mirror	misery	misfortune	mist	mode
model	moment	monk	monkey	monster	month	monument	mood	moral
mortal	mosquito	motion	motive	mountain	movement	mule	multitude	murderer
muscle	musket	musketeer	muskrat	myriad	mystery	nation	native	nature
necessity	neck	needle	neighbor	nerve	net	newspaper	niece	night
noble	noise	nose	notion	novel	nut	oar	oath	objection
obligation	observation	obstacle	occupation	odor	office	officer	official	one
operation	opinion	opponent	opportunity	orchard	organ	orphan	other	outline
owner	oyster	pace	page	pair	palace	paper	parent	parish
park	particular	partner	party	passage	passion	patch	path	patient
peak	pearl	peasant	peculiarity	people	perception	performance	period	person
petticoat	philosopher	physician	pig	pillar	pillow	pine	pipe	pirate
pistol	pit	plain	plane	planet	plank	plantation	plate	platform
plum	poet	pole	politician	pool	portrait	position	possession	possibility
post	potato	power	prayer	precaution	prejudice	preparation	price	priest
prince	princess	principle	prisoner	privilege	problem	product	professor	prophet
proportion	proposal	proposition	prospect	provision	pupil	purpose	pursuit	quality
quantity	quarter	rascal	ray	reader	reality	recollection	regiment	region
relation	relative	religion	remnant	representative	resolution	resource	responsibility	ribbon
rider	right	river	road	robber	robe	roof	room	rope
ruler	rumor	sailor	saint	savage	scene	scheme	scholar	science
scrap	scripture	sea	secret	section	senior	sensation	sentiment	sentinel
servant	service	sex	shadow	shaft	sheet	shilling	shirt	shoe
shop	shore	signature	silk	singer	sister	situation	skin	sky
sledge	sleeve	slice	slope	snake	soldier	son	song	sorrow
soul	space	spear	specimen	spectacle	spectator	speculation	speech	spirit
stable	stage	stair	star	statement	station	steamer	stone	story
stranger	street	string	stroke	structure	student	success	suggestion	summit
suspicion	sword	sympathy	system	table	tale	talent	tank	task
teacher	temple	temptation	tendency	tent	terror	theory	thing	thousand
threat	tiger	title	toe	tongue	torrent	tower	town	trader

Table 5: Unambiguous nouns (753) used in our experiments. Continuation of Table 4.

traveler	tree	trial	tribe	troop	truth	turkey	turnip	turtle
twig	twin	type	tyrant	valley	vapor	variety	vein	verse
vessel	vice	victim	victory	village	villain	vine	virtue	vision
visitor	voice	volume	wagon	wall	wand	war	warrior	way
weapon	week	well	wheel	window	wine	wing	wit	wood
word	worker	world	writer	yard	year			

Table 6: Unambiguous verbs (52) used in our experiments.

allow	appear	arise	ask	attract	become	begin	belong
bring	carry	come	contain	continue	depend	describe	deserve
do	eat	enter	exist	extend	follow	forward	give
grow	happen	hear	insist	involve	learn	occur	owe
own	perceive	possess	prevent	prove	put	receive	remember
remind	require	send	serve	shine	sit	speak	suggest
take	tell	tend	think				

Table 7: Full regression model fits for predicting rank outcomes from individual trial properties, following the regression model given in Equation (5).

Variable	Wav2Vec	Word
Intercept	0.53	0.10
allomorph_from=S × allomorph_to=S	-0.20	-0.08
allomorph_from=S	-0.06	0.06
allomorph_to=S	-0.30	0.06
from_freq	-0.38	-0.04
inflection_from=VBZ × allomorph_from=S × allomorph_to=S	-0.77	-0.01
inflection_from=VBZ × allomorph_from=S	0.46	-0.01
inflection_from=VBZ × allomorph_to=S	0.87	-0.03
inflection_from=VBZ × inflection_to=VBZ × allomorph_from=S × allomorph_to=S	0.46	-0.31
inflection_from=VBZ × inflection_to=VBZ × allomorph_from=S	1.41	0.25
inflection_from=VBZ × inflection_to=VBZ × allomorph_to=S	-0.71	0.07
inflection_from=VBZ × inflection_to=VBZ	-5.17	-0.13
inflection_from=VBZ	0.07	0.14
inflection_to=VBZ × allomorph_from=S × allomorph_to=S	0.48	0.06
inflection_to=VBZ × allomorph_from=S	-2.06	0.08
inflection_to=VBZ × allomorph_to=S	-1.57	-0.21
inflection_to=VBZ	7.53	0.05
to_freq	-1.05	0.01

Model	Noun	Verb
Wav2Vec	10.59	1.22
Word probe	1.94	0.135

Table 8: Mean rank outcomes of the same-word evaluation of Appendix B.4. These values are a soft lower bound for the results in Figure 3.

Hyperparameter	Value
Dimensionality	32
Margin m	0.37590
Learning rate	0.00108
Weight decay	0.00607

Table 9: Optimal hyperparameters for the word probe at the 8th layer of Wav2Vec Base, selected on a held-out development set in order to maximize discriminability of word categories.

Base	Base (IPA)	Consistent	Consistent (IPA)	Inconsistent	Inconsistent (IPA)
bay	/beɪ/	*bays*	/beɪz/	*base*	/beɪs/
decree	/dɪkri/	*decrees*	/dɪkriz/	*decrease*	/dɪkris/
den	/dɛn/	*dens*	/dɛnz/	*dense*	/dɛns/
dew	/du/	*dews*, *dues*	/duz/	*deuce*	/dus/
display	/dɪspleɪ/	*displays*	/dɪspleɪz/	*displace*	/dɪspleɪs/
fall	/fɔl/	*falls*	/fɔlz/	*false*	/fɔls/
fay	/feɪ/	*phase*	/feɪz/	*face*	/feɪs/
fear	/fɪr/	*fears*	/fɪrz/	*fierce*	/fɪrs/
flee	/fli/	*flees*	/fliz/	*fleece*	/flis/
for	/fɔr/	*fours*	/fɔrz/	*force*	/fɔrs/
gray	/greɪ/	*gray's*, *grays*, *graze*	/greɪz/	*grace*	/greɪs/
hahn	/hɑn/	*hahn's*	/hɑnz/	*hans*	/hɑns/
hen	/hɛn/	*hens*	/hɛnz/	*hence*	/hɛns/
her	/hɝ/	*hers*	/hɝz/	*hearse*	/hɝs/
how	/haʊ/	*how's*	/haʊz/	*house*	/haʊs/
in	/ɪn/	*inns*, *ins*	/ɪnz/	*ince*	/ɪns/
jew	/dʒu/	*jew's*, *jews*	/dʒuz/	*juice*	/dʒus/
joy	/dʒɔɪ/	*joys*	/dʒɔɪz/	*joyce*	/dʒɔɪs/
julia	/dʒuljʌ/	*julia's*	/dʒuljʌz/	*julius*	/dʒuljʌs/
knee	/ni/	*knees*	/niz/	*niece*	/nis/
knew	/nu/	*news*	/nuz/	*noose*	/nus/
law	/lɔ/	*laws*	/lɔz/	*los*	/lɔs/
lay	/leɪ/	*lays*	/leɪz/	*lace*	/leɪs/
may	/meɪ/	*maize*, *maze*	/meɪz/	*mace*	/meɪs/
one	/wʌn/	*one's*, *ones*	/wʌnz/	*once*	/wʌns/
pay	/peɪ/	*pays*	/peɪz/	*pace*	/peɪs/
peer	/pir/	*peers*	/pirz/	*pierce*	/pirs/
play	/pleɪ/	*plays*	/pleɪz/	*place*	/pleɪs/
ray	/reɪ/	*raise*, *rays*	/reɪz/	*race*	/reɪs/
river	/rɪvɝ/	*river's*, *rivers*	/rɪvɝz/	*reverse*	/rɪvɝs/
rye	/raɪ/	*rise*	/raɪz/	*rice*	/raɪs/
sin	/sɪn/	*sins*	/sɪnz/	*since*	/sɪns/
soar	/sɔr/	*sores*	/sɔrz/	*source*	/sɔrs/
spy	/spaɪ/	*spies*	/spaɪz/	*spice*	/spaɪs/
true	/tru/	*true's*	/truz/	*truce*	/trus/

Table 10: Materials used in the forced-choice experiment. Token word embeddings are retrieved based on transcribed phonetic form (IPA given here). Corresponding orthographic forms from the LibriSpeech corpus consistent with these phonetic transcriptions are given for convenience.