Encoder-decoder models for latent phonological representations of words

Cassandra L. Jacobs
University of California, Davis
clxjacobs@ucdavis.edu

Frédéric Mailhot
Autodesk Inc.
fred.mailhot@autodesk.com

Abstract

We use sequence-to-sequence networks trained on sequential phonetic encoding tasks to construct compositional phonological representations of words. We show that the output of an encoder network can predict the phonetic durations of American English words better than a number of alternative forms. We also show that the model's learned representations map onto existing measures of words' phonological structure (phonological neighborhood density and phonotactic probability).

1 Introduction

The representation of linguistic categories is a fundamental problem in (psycho)linguistics and natural language processing. The formation of complex representations from more basic components is relevant at all levels of linguistic representation, semantic, syntactic, and phonological. Finding good representations for words' phonological[1] structure is critical in psycholinguistics, where we wish to understand the phonological structure of the lexicon, which has been shown to be relevant for language comprehension and production.

The distributional hypothesis defines a word by the context in which it occurs (Harris, 1954; Firth, 1957). This approach has been extended more recently to other types of compositional structures, for example in characterizing the meanings and forms of sentences (Cer et al., 2018; Joulin et al., 2017; Conneau et al., 2017; Devlin et al., 2018). In this paper we explore whether distributional approaches can capture important phonological dependencies.

[1]There are disagreements in the literature about the location (Hale and Reiss, 2008) and even existence (Ohala, 1990b) of the boundary/interface between phonetics and phonology, so we remain as theory-agnostic as possible, freely using "phonological"/"phonetic" and "segment"/"phone" interchangeably.

Specifically, we test the extent to which recurrent *encoder-decoder* models (Cho et al., 2014; Sutskever et al., 2014) can learn representations that characterize the phonological structure of the lexicon while also having linguistic and psychological validity (Sibley et al., 2008). We propose that this approach can be used to learn viable lexical-level phonological representations. The output of the encoder component of our model yields promising results in the prediction of phonetic duration, outperforming a number of alternate phonological representations of words.

2 Quantifying a word's phonology

Given a set of discrete phonetic symbols i.e. graphemes with conventionalized pronunciations such as the International Phonetic Alphabet, it is trivial to represent any word's pronunciation as a sequence of such symbols. Conversely, relating sequences of such symbols (*viz.* words) to each other, as well as to the entire lexicon is less obvious. This challenge has led to a proliferation of measurements that characterize a word's phonetic or phonological relationship with all other words in the lexicon. We summarize some salient examples below, and briefly discuss some of their shortcomings.

2.1 Metrics insensitive to serial order

Phonological neighborhood density (PND). This measure is defined as the number of words having a Levenshtein edit distance of one from a given word (in terms of phonetic or phonological symbols) (Luce and Pisoni, 1998; Levenshtein, 1966). Under this definition, a word like "cat" has many neighbors, while a word like "molt" has fewer. This measure is simple to calculate and a wide variety of resources exist for obtaining these measures across many languages (Marian et al., 2012; Baayen et al., 1993; Luce and Pisoni, 1998).

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 206–217
Florence, Italy. August 2, 2019

While conceptually simple, PND is insensitive to the position of a segment within a word (e.g. word-initial versus word-final substitutions), and so "sat" and "cab" are treated as equally similar to "cat". Additionally, identifying a word's phonological neighbors using the Levenshtein distance metric requires specifying how many sounds can be added, deleted, or substituted, and potentially the allowable edit distance[2], increasing the number of choice points in determining what a "neighborhood" is.

Frequency-weighted phonological neighborhood density. An augmented version of PND, which weights phonological neighbors in proportion to their lexical frequencies (standardly estimated from large corpora; Marian et al., 2012). So, a more common word like "hat" would contribute more to the neighborhood density of "cat" than a less common word like "cap", even though they are at equal string edit distance. Whether and to what extent density measures should be frequency-weighted is an empirical question, though these measures seem to better reflect psycholinguistic processes than frequency-insensitive measures.

Feature-wise similarity. In the phonological literature it is standard to represent segments as collections of articulatory or acoustic features, e.g. [+voice], [-obstruent] (Chomsky (1968) is the canonical reference). Some linguists (e.g. Frisch (1996), *inter alia*) have posited that words like "cat" and "cap", which differ only in the place of articulation of their final segments (alveolar versus labial), should be considered more similar than e.g. "cat" and "can", which differ in both voicing and manner of articulation. This measure of similarity is potentially controversial, as there are theoretical and empirical questions as to which features to include, or even whether phonetic features exist at all (Stevens and Blumstein, 1981; Marslen-Wilson and Warren, 1994).

2.2 Metrics incorporating serial order

All of the previously described measures effectively characterize words as unordered collections of segments. These characterizations are incomplete because they fail to capture the fact that words unfold over time in usage. Representing the positions of phones within a word is critical for explaining a number of aspects of language processing. For example, the beginnings of words contribute more strongly than their ends to psycholinguistic effects that are attributed to their phonological representations (Levelt et al., 1999; Sevald and Dell, 1994, *inter alia*), and a word's phonological similarity to the rest of the words in the lexicon has important consequences for speech comprehension (Buz and Jaeger, 2016; Metsala, 1997). Some computational models encode segments as a function of their linear position within a syllable, e.g. in a *onset-vowel-coda* format (e.g. Dell, 1986; Sevald and Dell, 1994). Other approaches include segment n-grams to encode local aspects of serial order (e.g. Seidenberg and McClelland, 1989; Davis, 2010) and the oft-lamented Wickelphone (Houghton and Hartley, 1996). Most closely related to the present approach, some work has demonstrated the viability of sequence encoder models for representing sequences of characters or phonetic segments (Sibley et al., 2008).

2.3 Incorporating variability into representations

Psycholinguistic measures that quantify words' phonological properties in the lexicon generally ignore their variability in pronunciation. In usage, segmental context, or lexical factors such as word frequency, can significantly influence the phonetic realization of a given phone, ranging from assimilatory processes (Ohala, 1990a) to massive reduction and even complete omission (Pitt et al., 2005; Johnson, 2004, *inter alia*). For example, there are over 200 distinct transcriptions of the word "and" in the Buckeye corpus (Pitt et al., 2005), and its normative, dictionary pronunciation (i.e. [ænd]) only accounts for 3% of its realizations.

Measures such as PND rely on single, fixed pronunciations (generally normative/dictionary-based) and corpus-derived lexical frequencies to estimate how many similar-sounding words a given word has, but take no account of variability in realization. As there is evidence that listeners remember and can access/use individual exemplars of perceived speech (Pierrehumbert, 1980; Goldinger, 1998), it seems natural to model distinct realizations within the lexical network. The variability in a word's realizations may especially matter for identifying phonological competitors (Luce and Pisoni, 1998; Marian et al., 2012; Vaden et al., 2009). For example, words like "sand" and

[2]See e.g. Suárez et al., 2011 who allow edit distance greater than one, and track the mean distance to a fixed number of neighbors

"and" may rarely compete during lexical access, given that "and" is rarely pronounced similarly to "sand." By incorporating the variability available in naturalistic speech corpora, we hope to provide a better characterization of a word's phonological properties and its relation to the lexicon.

3 Latent phonological representations

Representing arbitrary-length sequences of phones with a single distributed representation has a number of potential practical and conceptual advantages. On the practical side, these representations have a fixed dimensionality, so finding meaningful groupings or clusters is computationally more tractable than directly clustering variable-length sequences. Moreover, projecting these sequences into a latent space offers the potential of discovering hidden relationships or variables that affect phonological or lexical structure.

Our aim in this paper is to test whether and to what extent recent approaches to building sentence representations can also be applied to the phonological domain. Both simpler and more complex latent representations can be constructed to characterize the phonological forms of words. We first discuss potential "naïve" means of accomplishing this, and then move into discussion of our proposed model.

Principal components on bag-of-n-phones

A number of document classification schemes and information retrieval tasks have treated documents as a product of the vector representations of words learned by principal components analysis (PCA; Landauer and Dumais, 1997). We apply this to the phonetic domain as well. By analogy to a bag of words, we refer to bag-of-phones (unigram features) and bag-of-n-phones (higher-order segment co-occurrence categories), which can then be fed into a dimensionality reduction algorithm like principal components analysis (PCA) as an approximate composition function to produce latent phonological representations of words.

doc2vec

Another dimensionality reduction method extends the continuous bag-of-words algorithm used to learn word vectors (Mikolov et al., 2013) to the document domain. Specifically, the model learns to compose (predict) a document (i.e. a word) from its phonological contents. doc2vec (Le and Mikolov, 2014) has been used in information retrieval and natural language processing applications (Lau and Baldwin, 2016) and so may be a viable way to obtain lexical phonological representations. As with bag-of-phones, this model is insensitive to serial order.

Sequential representations

Encoder-decoder or *sequence-to-sequence* (*seq2seq* henceforth) neural network architectures have shown considerable success in encoding sentences (*viz.* sequences of words) for tasks such as machine translation (Sutskever et al., 2014; Cho et al., 2014). These methods may be appropriate as a means of composing segmental representations, as they are intrinsically sensitive to ordering, easily take usage frequencies into account (directly from training corpora), and have been shown to be effective learners of sequential distributional properties of their training data.

4 Seq2seq model

We trained seq2seq models to either reproduce their input, or to recover (predict) normative (dictionary) pronunciations from the phonetic transcriptions of words in the Buckeye corpus (Pitt et al., 2005), a dataset of monologues provided in response to interviewer questions about the talkers' hometown of Columbus, Ohio. The corpus contains approximately 300,000 words.

Data inclusion criteria. There are some transcription errors in the Buckeye corpus, and so we excluded combinations of phones that did not occur at least ten times. This removes many errors, but a few remain. For example, the segment "h" occurs in some transcriptions but is not part of the character set of the transcription dictionary, and is thus likely an error of omission for actual digraphs from the dictionary; "th", "hh", etc. Despite the presence of these remaining errors, we do not correct the transcriptions of any words. In total, 57 phone/segment categories are represented. Full documentation of the coding scheme used in the corpus can be read in Pitt et al. (2005). For bag-of-n-phones features, we add the additional characters "w_s" and "w_e" as word boundary characters, signaling the starts and ends of words, respectively.

There are no standard train/dev/test splits for the Buckeye corpus, and so we restricted ourselves to randomly selected 80/20 train/test split (Pitt et al., 2005) for training all models.

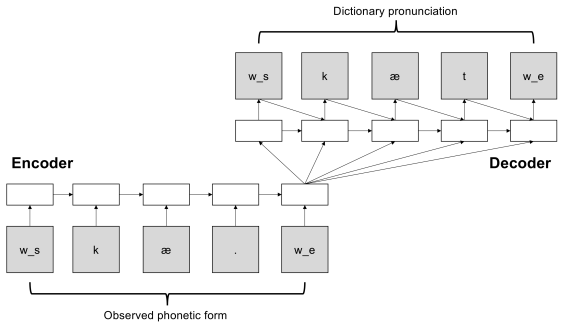

Figure 1: Encoder-decoder LSTM architecture (Normative decoder; for the Observed decoder, the output is the observed phonetic sequence).

Model architecture. Methodologically, we approach the problem with an eye to restricting the computational power of our model, and to restricting the space of hyperparameters to explore. To this end, our models use a basic recurrent encoder-decoder architecture, with an input-side embedding layer, and single-layer, unidirectional[3] LSTMs (Hochreiter and Schmidhuber, 1997) on the encoder and decoder sides. The encoder takes as input a sequence of phone indices (e.g. "cat" → `['k', 'ae', 'tq']` → [11, 1, 20]), embeds them, and encodes the sequence in the space defined by the LSTM. The encoder LSTM's final hidden state is provided as input to the decoder, whose task is to "unroll" this latent representation. The outputs of the decoder LSTM are successively fed through a softmax, sequentially outputting class probabilities for each character class in the phone vocabulary, which are then decoded via simple *argmax* (see Figure 1).

4.1 Training

Hyperparameters. The number of training epochs was empirically determined on the basis of asymptoting training loss, which we determined to be 25 epochs. We used a cross-entropy loss function, using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001. Other Adam parameters were at default values in the `dynet` python implementation as of this writing (version 2.0.3; Neubig et al., 2017). All hyperparameters were selected on the basis of asymptoting loss on a small subset of the training set. The embedding layer had 32 dimensions, and the encoder and decoder LSTMs were 64-dimensional.

Tasks. We trained two models to perform slightly different decoding tasks; the *Normative Decoder* model, and the *Observed Decoder* model. In both tasks, the inputs are transcriptions of observed realizations of words in the Buckeye corpus, which include e.g. phonetic changes and omissions. The *Normative Decoder's* task is to output the word's normative pronunciation (e.g. `[k, ae, tq]` → `[k, ae, t]`), while the *Observed Decoder* model is trained as a sequential autoencoder (e.g. Chung et al., 2016); the task is to reproduce the input sequence exactly. Both are potentially viable approaches to the creation of lexical phonological representations and show similar performance in the downstream tasks reported on below, which may be useful for researchers who only have access to normative pronunciations.

We evaluated the performance of the model on the 20% held-out portion of the corpus.

4.2 Lexical representations

Once the model is trained, any sequence of phones can be input to the encoder, yielding a latent phonological representation of that sequence. As with character-based NLP models, the comparatively low dimensionality of the input space (57 segments) mitigates sparsity issues, consequently we can obtain latent phonological representations not just of vocabulary words that have been trained but also for rare, out-of-vocabulary (OOV) words and non-words. We plot some aspects of the learned representations in Figures 2 and 3. One pattern that is particularly apparent is that the left-to-right serial nature of the encoder leads to representations that strongly encode the final segment in their representations, for both consonants and vowels.

5 Evaluation

As a preliminary investigation of the information encoded in the learned lexical representations, we assess their ability to model phonetic duration, which is known to be sensitive to phonotactic probability and phonological overlap (Gahl et al., 2012; Watson et al., 2015; Buz and Jaeger, 2016; Yiu and Watson, 2015; Goldrick and Larson, 2008; Vitevitch and Luce, 2005), in addition to other factors like contextual predictability (e.g.

[3] While we do not perform these experiments here, we believe that a Bi-LSTM encoder (Schuster and Paliwal, 1997) will enable further advances in constructing psycholinguistically predictive word representations.

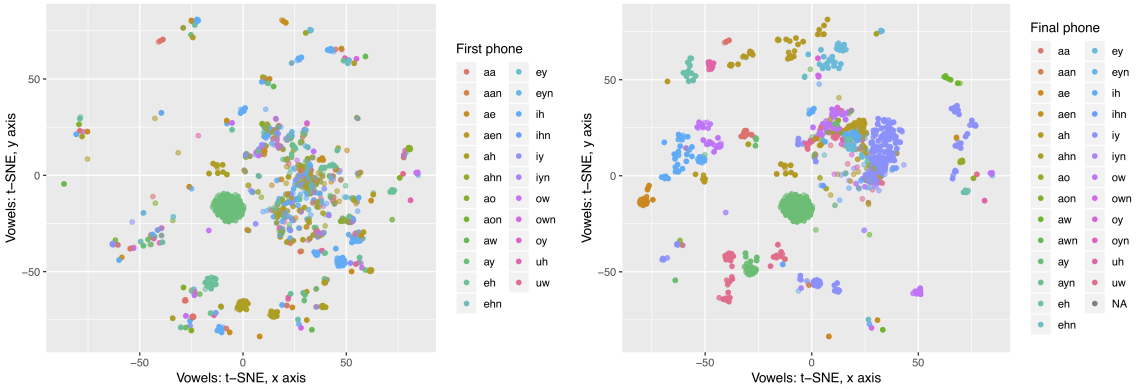

Figure 2: Topology of word vectors from phonological encoder models learned by t-SNE (Maaten and Hinton, 2008). Degree to which word vectors encode vowel information. Clusters largely prioritize word-final information, especially the last segment. Left graph represents the identities of the first segment. Right graph represents the identities of the final segment. The strong encoding of the final segment may be due to the model architecture using uni-directional recurrent layers.

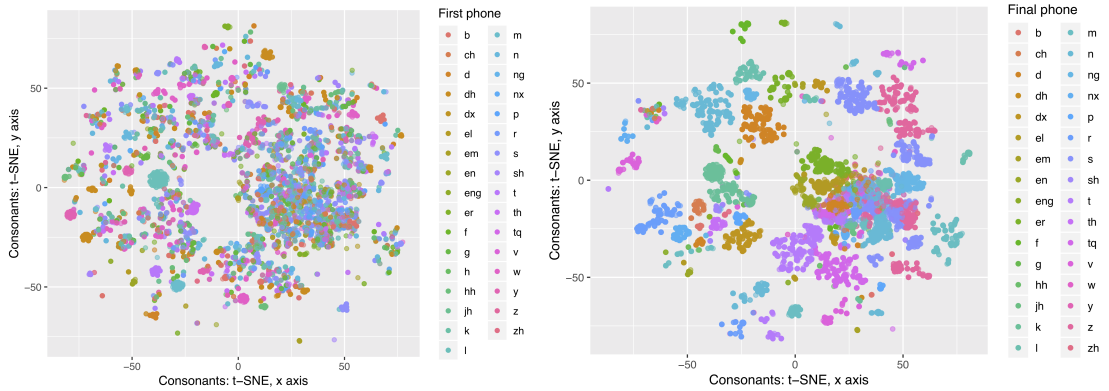

Figure 3: Topology of word vectors, t-SNE projection (Maaten and Hinton, 2008). Degree to which word vectors encode consonant information. Clusters largely prioritize word-final information, especially the last segment. Left graph represents the identities of the first segment. Right graph represents the identities of the final segment.

Cohen Priva and Jaeger, 2018; Seyfarth, 2014). We show that the encoder creates sequence representations that are useful for predicting word duration, and compare the success of the encoder to several other models, described below.

5.1 Predicting word duration

Ultimately we are interested in whether latent phonological representations have predictive validity for phonetic cues, potentially in conjunction with other phonological and lexical representations. Word duration has been shown to be strongly related to phonological structure (Gahl et al., 2012), because duration may reflect the mechanics of the phonological sequencing process in language production (Yiu and Watson, 2015; Watson et al., 2015; Fox et al., 2015) or because speakers lengthen words in dense neighborhoods to promote the listener's understanding (Tily and Kuperman, 2012).

We built a series of nested statistical models designed to predict whole-word phonetic duration. The durations were obtained by summing up the durations of each of the annotated phonetic segments for an individual word, which are themselves derived from time stamps extracted from the Buckeye metadata. Whole-word durations were log transformed due to their positive skew; failing to account for this can make statistical inference more difficult (Campbell, 1992). All models were constructed using ridge (L1 norm) regression using the `scikit-learn` package in Python (version 0.2.0; Pedregosa et al., 2011). We report goodness of fit measures in all cases by R^2 values (the coefficient of determination; provided automatically by the *score* function within the ridge regression model object).

All duration models were trained on the same

80-20% split that was used to train the encoder-decoder. Consequently, there were 282,742 observations (words) during training, and 70,686 words at test. The vocabulary for the bag-of-words representations was estimated from the training data. All models are summarized in Table 1.

5.2 Baseline models

Word embeddings. A word's distributional properties, such as its part of speech and meaning; latent part-of-speech; or word-frequency information may reliably predict a word's duration (Seyfarth, 2014; Turnbull et al., 2018; Priva, 2015). Consequently, we incorporate 100-dimensional word embeddings into the regression models. We obtained these word embeddings from gensim's (Řehůřek and Sojka, 2010) skip-gram implementation trained on the Fisher corpus (Cieri et al., 2004), which we selected due to its size, which is critical for generating good word embeddings (Antoniak and Mimno, 2018), and because it belongs to the same domain as the Buckeye corpus (conversational speech).

The skip-gram model used a context window of 5 words and a negative sampling size of 5. We used a zero vector to represent OOV (e.g. Columbus, Ohio-specific place names that would not occur in the Fisher corpus). Word embeddings were, on their own, not a strong predictor of word duration ($R^2 = 0.082$) on the test set, but nevertheless account for some of the variance in word duration.

Bag-of-phones models. Bag-of-words representations are a useful and informative baseline in other NLP tasks, especially text classification (Wang and Manning, 2012). We obtained bag-of-phone representations by learning a vocabulary on the training data and creating sparse count vectors in which the features represent individual phones. A simple bag-of-uniphones model, which ignores order information, has greater predictive power than word embeddings on the test set (R^2=0.140). This shows that it is possible to at least partly predict the duration of a given word's realization from relatively unstructured phonological information.

Bag-of-n-phones. Unlike bag-of-words representations, bag-of-ngrams encode localized order information. We constructed n-gram features of phone combinations (bag-of-n-phones) of lengths 2 to 5, using a cutoff frequency of 10 observations. These more complex representations performed similarly to the simpler bag-of-phones model on the test set ($R^2 = 0.140$).

We also tested whether incorporating word boundary information into these models ("w_s" and "w_e" phones) would induce boundary-sensitive phonotactics, but this also did not provide additional gains over simpler models (R^2 = 0.138 and R^2 = 0.140).

Principal components analysis over bag-of-n-phones. Following from the previous section, we take our bag-of-n-phones representations and feed them into a truncated singular value decomposition model to obtain latent representations of words ("documents"). This representation explained a slightly greater amount of variance in word duration than word embeddings (R^2 = 0.106). However, this method performed far worse than the bag-of-phones and bag-of-n-phones models described in the previous section, indicating that some information is lost in this dimensionality reduction method.

doc2vec. Our doc2vec model vectors were trained to predict a word from a phonological representation. The resulting vectors had the same dimensionality as the PCA vectors and the encoder output of the seq2seq models. Surprisingly, doc2vec performed the worst of models that we considered (R^2 = -0.05).

seq2seq. The outputs of the encoders for the Observed and Normative decoder models were among the best we considered, both on their own and in conjunction with other measures. Interestingly, the *Observed Decoder* provides a much closer fit to phonetic duration than word embeddings, bag-of-phones, PCA, doc2vec, and the *Normative Decoder* representations. When combined with bag-of-phones and word embedding information, the *Observed Decoder* representations explain the greatest amount of variance in word duration (R^2 = 0.181), suggesting that these latent phonological representations encode useful information for characterizing word form.

The disparity between the *Observed* and *Normative* decoder models may be a consequence of the *Normative* model's more difficult learning problem. One potential explanation is that despite training the two models for equal lengths of time (25 epochs), the Normative decoder was not trained to the same criterion as the Observed decoder. Future work should explore whether the worse performance of the Normative decoder model is due to the precision of its representations

Simple	Test R^2	No. features	**Combined**	Test R^2	No. features
Word embeddings (WE)	0.082	100	**BoP + wb + WE**	0.161	159
Bag-of-phones (BoP)	0.140	57	**+ Observed decoder**	**0.181**	223
+ w_s + w_e (wb)	0.140	59	+ Normative decoder	0.177	223
Bag-of-n-phones (BoNP)	0.140	1700	BoNP + wb + WE	0.159	5018
+ w_s + w_e (wb)	0.138	4918	+ Observed decoder	0.175	5082
PCA bag-of-n-phones	0.106	64	+ Normative decoder	0.173	5082
doc2vec	-0.05	64	Observed + WE	0.149	164
Observed decoder	**0.149**	64	Normative + WE	0.141	164
Normative decoder	0.140	64			

Table 1: Ablation study. Effectiveness of features and combinations of features for predicting (log) phonetic duration.

or due to what is embedded in the representations themselves.

6 Probing phonological structure

While it is clear that seq2seq representations of the phonological forms of words are partially predictive of a phonetic phenomenon (duration), whether the representations encode anything useful about the lexicon requires further investigation. In this section, we explore whether characterizing the similarity space of these phonological word vectors can approximate standard measures of a word's phonological properties. The results show that the vectors produce coherent clusters of words with different phonological properties. We also show that there are correlations between our measures and phonotactic probability.

6.1 Latent phonological neighborhood density

While it is not commonly the case that similarity scores follow a normal distribution, in our case, the similarity scores for words are by visual spot inspection roughly symmetric and normally distributed, so we chose to characterize individual words w_i by the mean and standard deviation of their similarity scores to every other word in the lexicon. Although not *a priori* obvious, one possibility is that these metrics correlate with other lexical metrics, for example, a wide standard deviation could mean that a word has a number of different ways it can be similar to other words, whereas a narrow standard deviation suggests that the word is fairly unique.

6.2 The similarity structure of the lexicon

The distributions of similarity scores show some interesting properties. Unlike the measurements of phonological neighborhood density provided in Vaden et al. (2009), which follow a quasi-Zipfian distribution, a histogram of the mean word-lexicon similarities across the whole vocabulary shows a very different pattern. In particular, there appear to be three distinct clusters of similarity scores, as shown in Figure 4.

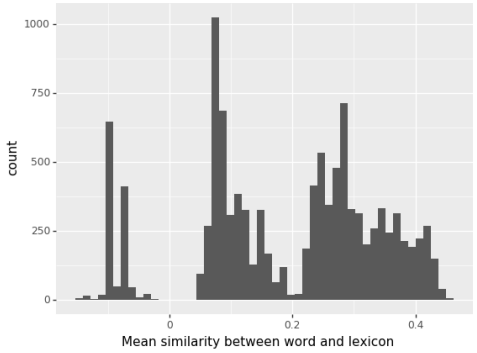

Figure 4: Three clusters of similarity scores from Observed Decoder model.

Words in the first cluster, which show negative average similarity scores, were highly frequent words, typically encompassing function words (e.g. *but*, *about*, *the*). The second cluster appeared to include less high-frequency terms (e.g. *day*, *brain*, *wants*). Finally, the rightmost cluster typically had higher similarity scores, representing low frequency and longer words (e.g. *devices*, *widely*, *element*).[4] Going forward, a meta-model

[4]We thank our reviewers for pointing out that all of these properties are correlated with word length in segments (e.g. highly frequent words are on average shorter), which is a useful baseline that we will explore in future work.

will be necessary to determine what factors determine a word's mean lexicon-similarity value.

6.3 Correlation with existing phonological properties

Ideally, a new measure of phonological form should relate to measures already known to affect speech production. For example, a significant correlation with a particular word's mean or standard deviation similarity to all the other words in the lexicon would suggest that our measures characterize the lexicon in a similar way to existing measures. Similarly, because our latent representations encode sequences, we expect them to correlate with phonotactic probability (Vitevitch and Luce, 2004). So, as a final set of analyses, we sought to test whether and to what extent the Observed decoder learns representations that can tell us about a word's relationship to the rest of the lexicon.

There are two measures of interest that have received some attention in the speech production literature. For the present analyses, we reference the phonological neighborhood density metrics as well as the phonotactic probability scores for words in Buckeye that are also in the Irvine Phonotactic Online Dictionary (IPhOD; Vaden et al., 2009). We show that our measures (both mean and standard deviation) strongly correlate with phonotactic probability and IPhoD's additional PND measure. This suggests that the vectors' usefulness extends to researchers who wish to explore the phonological similarity structure of the lexicon for psycholinguistic research.

Phonological neighborhood density. Given the importance of phonological neighborhood density (PND) in speech production (Luce and Pisoni, 1998; Vitevitch and Luce, 2005; Metsala, 1997; Mirman, 2011), we correlated the (log) number of phonological neighbors with our latent density scores and phonetic duration. A phonological neighbor is a word that differs by a single sound (either an addition, a substitution, or a deletion; Levenshtein, 1966). PND ((log) # of neighbors, Figure 5) has a strong negative correlation with mean word-lexicon similarity (greater mean similarity translates to fewer neighbors; ρ = -.59) while the standard deviation of word-lexicon similarity shows a non-linear relationship with neighborhood density.

Phonotactic probability. Phonotactic probability is a measure of the phonological typicality of a word, computed from product of uni-phone and bi-phone probabilities of that word pronunciation, in the same fashion that sentence probabilities are computed in a standard bigram language model (Vitevitch and Luce, 2004, 2005). In our final analysis, we compare the mean and standard deviation of a word's similarity to all other word types, including alternate pronunciations of the same word, to existing measures of phonotactic probability. As with phonological neighborhood density, we see significant positive correlations between our phonological similarity measures (both means and standard deviations; ρ = 0.41 and ρ = 0.13, respectively) between phonotactic probabilities, which we visualize in Figure 5.

7 Conclusion

The results presented here suggest that encoder-decoder models are a promising framework for composing segment-based representations of words. The models also characterize words' phonological forms relative to the rest of the lexicon. We believe that encoder-decoder models' usefulness extends beyond that of many existing approaches, as they can seamlessly generate gestalt representations for out-of-vocabulary words and even nonce words. Our approach has a number of potential advantages for the cognitive modeling of language processing in both comprehension and production tasks, or indeed in any task that can be modeled with phonological word representations. Importantly, the encoder-decoder modeling framework is flexible, learning both from observed, quasi-phonetic realizations of words as well as from idealized, normative (dictionary-based) pronunciations, and allows for many variations in expressivity and computational power.

The reported correlations between phonological neighborhood density, phonotactic probability, latent phonological similarity, and phonetic duration motivate a need to better understand the embedding representations themselves. We have presented considerable evidence that the models capture some non-trivial dependencies between phonetic segments that can characterize word forms. Going forward, we believe that our latent phonological representations may be useful for designing stimuli, or provide an alternative to standard covariates in psycholinguistic experiments such

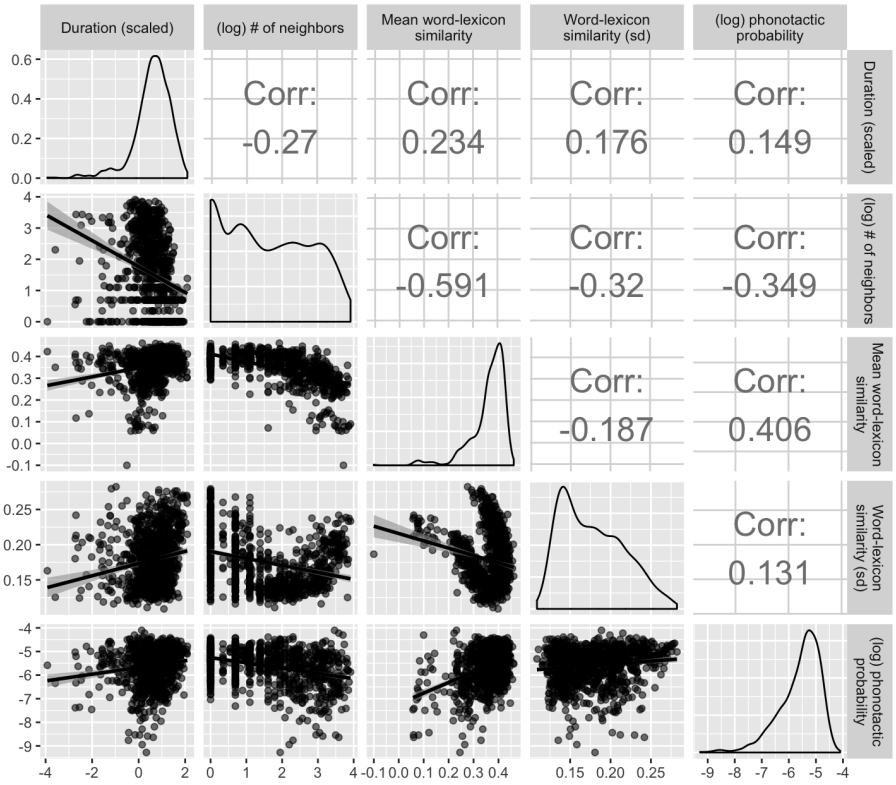

Figure 5: Correlation between a word's phonetic duration in Buckeye, phonological neighborhood density, global word-lexicon similarity (mean and standard deviation), and phonotactic probability.

as phonological neighborhood density and phonotactic probability. Finally, our results on the Normative-Decoder suggest that low-resource languages with only a pronunciation dictionary are also a viable means of learning these representations, assuming that there is a corresponding corpus of conversational data. In sum, we have demonstrated that our approach is useful for modeling of phonological structure.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association of Computational Linguistics*, 6:107–119.

R Harald Baayen, Richard Piepenbrock, and Rijn van H. 1993. The {CELEX} lexical data base on {CD-ROM}. *Linguistic Data Consortium*.

Esteban Buz and T Florian Jaeger. 2016. The (in) dependence of articulation and lexical planning during isolated word production. *Language, Cognition and Neuroscience*, 31:404–424.

W Nick Campbell. 1992. Syllable-based segmental duration. *Talking machines: Theories, models, and designs*, pages 211–224.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Noam Chomsky. 1968. *The sound pattern of English*. Studies in language. Harper & Row, New York.

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *Interspeech 2016*, pages 765–769.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Uriel Cohen Priva and T Florian Jaeger. 2018. The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard*, 4.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Colin J Davis. 2010. The spatial coding model of visual word identification. *Psychological Review*, 117:713–758.

Gary S Dell. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93:283–321.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Neal P Fox, Megan Reilly, and Sheila E Blumstein. 2015. Phonological neighborhood competition affects spoken word production irrespective of sentential context. *Journal of Memory and Language*, 83:97–117.

Stefan Frisch. 1996. *Similarity and frequency in phonology*. Ph.D. thesis, Northwestern University.

Susanne Gahl, Yao Yao, and Keith Johnson. 2012. Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66:789–806.

Stephen D Goldinger. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105:251–279.

Matthew Goldrick and Meredith Larson. 2008. Phonotactic probability influences speech production. *Cognition*, 107:1155–1164.

Mark Hale and Charles Reiss. 2008. *The Phonological Enterprise*. Studies in language. Oxford University Press, New York.

Zellig S Harris. 1954. Distributional structure. *Word*, 10:146–162.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

George Houghton and Tom Hartley. 1996. Parallel models of serial behaviour: Lashley revisited. *Psyche: An Interdisciplinary Journal of Research on Consciousness*.

Keith Johnson. 2004. Massive reduction in conversational american english. In *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*, pages 29–54. Citeseer.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 427–431.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–38.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Paul A Luce and David B Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19:1–36.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, 7(8):e43230.

William Marslen-Wilson and Paul Warren. 1994. Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological review*, 101(4):653.

Jamie L Metsala. 1997. An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition*, 25(1):47–56.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Daniel Mirman. 2011. Effects of near and distant semantic neighbors on word production. *Cognitive, Affective, & Behavioral Neuroscience*, 11(1):32–43.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

John J Ohala. 1990a. The phonetics and phonology of aspects of assimilation. *Papers in Laboratory Phonology*, 1:258–275.

John J Ohala. 1990b. There is no interface between phonology and phonetics: a personal view. *Journal of Phonetics*, 18:153–171.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Janet Breckenridge Pierrehumbert. 1980. *The phonology and phonetics of English intonation*. Ph.D. thesis, Massachusetts Institute of Technology.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45:89–95.

Uriel Cohen Priva. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2):243–278.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

M. Schuster and K.K. Paliwal. 1997. There is no interface between phonology and phonetics: a personal view. *IEEE Transactions on Signal Processing*, 45:2673–2681.

Mark S Seidenberg and James L McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:523–568.

Christine A Sevald and Gary S Dell. 1994. The sequential cuing effect in speech production. *Cognition*, 53:91–127.

Scott Seyfarth. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Daragh E Sibley, Christopher T Kello, David C Plaut, and Jeffrey L Elman. 2008. Large-scale modeling of wordform learning and representation. *Cognitive Science*, 32(4):741–754.

Kenneth N Stevens and Sheila E Blumstein. 1981. The search for invariant acoustic correlates of phonetic features. *Perspectives on the study of speech*, pages 1–38.

Lidia Suárez, Seok Hui Tan, Melvin J Yap, and Winston D Goh. 2011. Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, 18(3):605–611.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Harry Tily and Victor Kuperman. 2012. Rational phonological lengthening in spoken dutch. *The Journal of the Acoustical Society of America*, 132(6):3935–3940.

Rory Turnbull, Scott Seyfarth, Elizabeth Hume, and T Florian Jaeger. 2018. Nasal place assimilation trades off inferrability of both target and trigger words. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 9(1).

Kenneth I Vaden, HR Halpin, and Gregory S Hickok. 2009. Irvine phonotactic online dictionary, version 2.0. [data file]. *Available from http://www.iphod.com.*

Michael S Vitevitch and Paul A Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, 36:481–487.

Michael S Vitevitch and Paul A Luce. 2005. Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, 52:193–204.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: simple, good sentiment and topic classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, 2:90–94.

Duane G Watson, Andrés Buxó-Lugo, and Dominique C Simmons. 2015. The effect of phonological encoding on word duration: Selection takes time. In *Explicit and implicit prosody in sentence processing*, pages 85–98. Springer.

Loretta K Yiu and Duane G Watson. 2015. When overlap leads to competition: Effects of phonological encoding on word duration. *Psychonomic Bulletin & Review*, 22:1701–1708.