2018

Sound Analogies with Phoneme Embeddings

Miikka P. Silfverberg  
*University of Colorado*, miikka.silfverberg@colorado.edu

Lingshuang Mao  
*University of Colorado*, lima4664@colorado.edu

Mans Hulden  
*University of Colorado*, mans.hulden@colorado.edu

Follow this and additional works at: [http://scholarworks.umass.edu/scil](http://scholarworks.umass.edu/scil)

Part of the [Computational Linguistics Commons](http://scholarworks.umass.edu/scil)

**Recommended Citation**

DOI: 10.7275/R5NZ85VD  
Available at: [http://scholarworks.umass.edu/scil/vol1/iss1/15](http://scholarworks.umass.edu/scil/vol1/iss1/15)

This Paper is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Sound Analogies with Phoneme Embeddings

Miikka Silfverberg and Lingshuang Jack Mao and Mans Hulden
Department of Linguistics
University of Colorado
first.last@colorado.edu

Abstract

Vector space models of words in NLP—word embeddings—have been recently shown to reliably encode semantic information, offering capabilities such as solving proportional analogy tasks such as man:woman::king:queen. We study how well these distributional properties carry over to similarly learned phoneme embeddings, and whether phoneme vector spaces align with articulatory distinctive features, using several methods of obtaining such continuous-space representations. We demonstrate a statistically significant correlation between distinctive feature spaces and vector spaces learned with word-context PPMI+SVD and word2vec, showing that many distinctive feature contrasts are implicitly present in phoneme distributions. Furthermore, these distributed representations allow us to solve proportional analogy tasks with phonemes, such as \textit{p} is to \textit{b} as \textit{t} is to \textit{X}, where the solution is that \textit{X} = \textit{d}. This effect is even stronger when a supervision signal is added where we extract phoneme representations from the embedding layer of an recurrent neural network that is trained to solve a word inflection task, i.e. a model that is made aware of word relatedness.

1 Introduction

Distributional word representations, or word embeddings, have attracted much attention in NLP, and their success is considered a vindication of the distributional hypothesis for lexical semantics espoused much earlier by the likes of Wittgenstein,\footnote{“the meaning of a word is its use in the language” \cite{Wittgenstein1953}} Firth,\footnote{“The complete meaning of a word is always contextual” \cite{Firth1937}} Harris,\footnote{“distributional statements can cover all of the material of a language” \cite{Harris1954}} and other contemporaries. Often overlooked is that this hypothesis among linguists has extended itself much wider to include phonology and grammar: “all elements of speech (phonological, lexical, and grammatical) are now to be defined and classified in terms of their relations to one another” \cite{Haas1954}.

Given the successes of distributional models not only in specifying semantic similarity, but also addressing proportional analogy tasks \cite{Turney2010, Mikolov2013a, Mikolov2013b, Levy2014}, we want to investigate if distributional representations of phonemes induce a similarly coherent space as lexical items do, and if the properties of such spaces conform to linguistic expectations, a question posed in another form as early as in Fischer-Jørgensen \cite{Fischer-Jorgensen1952}. In particular, we address two questions: (1) whether learned vector representations of phonemes are congruent with commonly assumed binary phonological distinctive feature spaces, and (2) whether a proportional analogy of the type \textit{a}:\textit{b}::\textit{c}:\textit{d} (\textit{a} is to \textit{b} as \textit{c} is to \textit{d}) discovered in a phoneme embedding space is also a valid analogy in a phonological distinctive feature space, where phonemes are represented in a space of standard articulatory binary features \cite{Mielke2008} such as ±continuant, ±voice, ±high, ±coronal, etc.

We address these questions using three different methods of obtaining vector representations of phonemes; two unsupervised models that associate phonemes and the contexts (neighboring phonemes) they occur in (word2vec, PPMI+SVD), and one model where we are given lemmas and inflected forms as supervised data for a recurrent neural network (RNN) encoder-
decoder model that is trained to perform such inflections from where we extract the vector representations in the embedding layer. The latter method has a weak supervision signal in that the system knows which words are truly related forms, since it is trained only on word pairs where one of the words is a result of transforming the other one according to inflectional features. Hence, this latter model can presumably better pick up on phonological alternations along distinctive feature lines that occur between phonemes; for example, knowledge of relatedness helps in determining that two forms exhibit a voicing alternation, as in the Finnish pair mato (‘worm’, nominative) ∼ madon (‘worm’, genitive), and hence, the segments t and d should align themselves in the embedding space parallel to other voiceless/voiced pairs that alternate similarly. We perform experiments in three languages: Finnish, Turkish, and Spanish.

2 Related Work

Local co-occurrence of phonemes and (in writing) phonemic graphemes has been widely explored in the literature for unsupervised discovery of phonological features. The observation of Markov (1913, 2006) that vowels and consonants tend to alternate in a statistically robust way in phonemic writing systems is an early observation that some articulatory features can be recovered in an unsupervised way. Algorithms for cryptographic decipherment often take advantage of such patterns (Guy, 1991; Sukhotin, 1962, 1973). Local co-occurrence counts have also been analyzed through spectral methods, such as singular value decomposition (Moler and Morrison, 1983; Goldsmith and Xanthos, 2009; Thaine and Penn, 2017), revealing that significant latent structure can be recovered, mainly with respect to vowels and consonants. Recent works along the same lines of inquiry include Kim and Snyder (2013) that presents a Bayesian approach that simultaneously clusters languages and reveals consonant/vowel/nasal distinctions in an unsupervised manner. Hulden (2017) shows that an algorithm based on the obligatory contour principle (Leben, 1973) and an additional assumption of phonological tiers being present (Goldsmith, 1976) robustly reveals at least consonant/vowel, coronal/non-coronal, and front/back distinctions from unlabeled phonetic data or orthographic data from phonemic writing systems.

Learning features directly from waveform representations (see e.g. Lin (2005))—while not addressed in this paper—is also highly relevant to the current study, and is indeed a question to which some lower-level, speech signal-based form of distributed representations may be adapted.

The idea explored in this paper—that phonemes (or graphemes) might exhibit linguistically apt correlations in an embedding space—has been implied by earlier research, for example Faruqui et al. (2016). In that work, a neural encoder-decoder model (Cho et al., 2014; Sutskever et al., 2014) was trained to perform a transformation of words from their citation forms to a ‘target’ inflected form and, after training, the vowels in the embedding layer of the long-short term memory (LSTM) neural model trained for Finnish were found to clearly group themselves according to known harmony patterns in the language. Li et al. (2016) take advantage of phoneme transcriptions in a neural speech synthesis application, showing improvements on this task and indicating that similar phonemes in a bidirectional LSTM (Bi-LSTM) embedding layer map close to each other. More closely related to the current work, Dunbar et al. (2015) investigate how well phonetic feature representations in English align with vector representations learned from local contexts of sound occurrence using both a neural language model and also a matrix factorization model. Their (surprisingly) negative results can be explained by the fact that the experiment was set up to compare two spaces with respect to all and only minimally differing pairs (such as: is the distinctive feature vector offset from p to b like the offset from t to d in the embedding space?) using a small specific fixed number of phonological features. By contrast, in our experiments, we learn a vector space model and examine if all phoneme-pair distances correlate globally between the embedding space and a known phonological distinctive feature space, and also whether analogies deemed to be ‘good’ in the embedding spaces are also ‘good’ in the distinctive feature space. This is less sensitive to an assumption that all distinctive features have correlates in the embedding space. For example, Figure 3 shows a space induced by one of our methods, where the offset from a to æ (low vowels) is similar to that of o to ø (high vowels), both being back-to-front vowel transformations, but where the equivalent harmonic corre-
3 Models and methods

We consider three different models for learning phoneme embeddings.

PPMI+SVD These embeddings are formulated using truncated Singular Value Decomposition (SVD) on a matrix of positive point-wise mutual information (PPMI) values (Bullinaria and Levy, 2007; Levy and Goldberg, 2014). For the definition of PPMI, see Equation 1. We first compute PPMI values for co-occurrences of center phonemes and context phonemes in a five character sliding window over the training data. We then arrange the PPMI values into an $n \times n$ matrix $M$ and apply SVD to get the factorization $M = UV^\top$, where $U$ and $V$ are orthonormal, and $\Sigma$ is diagonal. Let $U_d$ denote the $n \times d$ matrix derived from $U \Sigma$ by truncating all rows to length $d$. Then our $d$-dimensional phoneme embeddings are the rows of $U_d$.

$$\text{PPMI}(x, y) = \max(\log \frac{p(x, y)}{p(x)p(y)}, 0) \quad (1)$$

The probabilities $p(x, y), p(x), \text{and } p(y)$ are derived from simple counts by maximum likelihood estimation.

word2vec Our second model is the word2vec model introduced by Mikolov et al. (2013a) for modeling semantic relatedness of words. Like the PPMI+SVD model, the word2vec model captures distributional information about phonemes. However, it explicitly constructs a language model and trains embeddings which perform well on the language modeling task. Nevertheless, Levy et al. (2014) show that the models are in a sense the same: like PPMI+SVD, the skip-gram variant of word2vec with negative sampling is also implicitly performing a factorization of a matrix of shifted PMI values of words and their context words. Although the models are similar, we decided to include word2vec because of claims that it sometimes tends to give better results on analogy tasks (Levy et al., 2014).

In our experiments, we use standard word2vec embeddings generated with the gensim toolkit. We use the skipgram model and negative sampling with window size 1.

RNN encoder-decoder Our final model differs from the first two in that it learns embeddings which maximize performance on a word inflection task: the system receives lemmas and the morphological features of the desired inflected form as input and emits corresponding inflected forms. We formulate the system as an RNN encoder-decoder (Cho et al., 2014). Our system is identical to the system presented in Kann and Schütze (2016) except that it does not incorporate attention. The encoder is realized using a bidirectional LSTM model which operates on character/phoneme embeddings (see Figure 1). The same embeddings are also used by the decoder as explained more thoroughly in Kann and Schütze (2016). We first train the system and then extract the phoneme embeddings (see Figure 2) and use them as our phoneme vector representations.

The performance of RNN encoder-decoder models is known to be sensitive to the random initialization of parameters during training. In all ex-
Figure 2: A neural network learns to map a one-hot input into an intermediate representation (the embedding layer). This transformation is tuned to perform well on an inflection task and yields a dense vector representation of segments.

Experiments, we therefore train five separate models using five different random initializations of parameters and compute similarity scores and analogy scores as averages of the scores given by individual models.

4 Data and Resources

We train all models using Finnish, Spanish and Turkish data sets from the SIGMORPHON 2016 shared task for morphological re-inflection (Cotterell et al., 2016). Each line in the data sets contains an inflected word form, its associated lemma and morphological features. For Finnish, the training data consists of 12,692 lines, for Spanish, 12,575 lines and, for Turkish, 12,336 lines. We learn embeddings for orthographic symbols occurring more than 100 times in the respective data sets. For Finnish, this set includes 25 symbols, for Spanish, 28 symbols and, for Turkish, 27 symbols.

The PPMI+SVD and word2vec models only use word forms for training. In contrast, the RNN encoder-decoder is trained on all parts of the training set: word forms, lemmas and morphological features. For all three languages, we use the training data for subtask 1 of the shared task.

There is a near one-to-one correspondence between Finnish and Turkish graphemes and phonemes. For Spanish, the correspondence between the orthographic and phonetic representation of the language is, however, less straightforward. We therefore perform a number of transformations on the training data in order to bring it closer to a phonetic representation of the language. Specifically, we transform voiced stops b, d and g to the voiced fricatives with the same place of articulation postvocally (β, ð, γ). We add-

![Figure 3: A vowel space for Finnish learned by collecting local phone(me)-context counts (window size 5), followed by a PPMI transform, followed by SVD, truncated to three dimensions. Neutral vowels are e, i (circled in yellow) and the harmony correspondences are a (=IPA a) ∼ ā (= IPA æ), o ∼ ō (= IPA ø) and u ∼ y. The front harmonic group is circled in blue.](image)

Table 1: Features used in manually crafted articulatory representations.

<table>
<thead>
<tr>
<th>VOWELS</th>
<th>(Syllabic), Front, Back, High, Low, Round, Tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSONANTS</td>
<td>Consonantal, Sonorant, (Syllabic), Voice, Labial, Coronal, Dorsal, Pharyngeal, Lateral, Nasal, Continuant, Delayed Release, Distributed, Tap, Anterior, Strident</td>
</tr>
</tbody>
</table>

We compare embeddings discovered by different systems to manually crafted articulatory representations of phonemes/allophones based on standard IPA descriptions in Hayes (2011). The list of the phonetic features we use is given in Table 4. We realize the representations as vectors $v \in \{0, 1\}^n$ in a distinctive feature space, where $n$ is the number of distinctive features in the description (22 in our model). Each dimension in feature space corresponds to a phonetic feature such as continuant, syllabic and voice. Entry $i$ in a feature vector is 1 if the corresponding phoneme is positive for the given feature. Otherwise, it is 0.

5Our code is available at s://github.com/mpsilfvehttp/phonembedding
Table 2: Correlation between feature similarities $\text{sim}(\text{feat}(x), \text{feat}(y))$ and embedding similarities $\text{sim}(\text{emb}(x), \text{emb}(y))$ for all unordered pairs of phonemes \{x, y\} (where $x \neq y$). All correlations are significantly higher (with $p$-value < 0.01) than ones obtained using a random assignment of embedding vectors to phonemes.

Table 3: The embedding space is used to generate an $n$-best list of ab:c:d analogy proposals. The table shows the average number of differing distinctive features between $d$ and $X$ when $X$ is calculated by the same analogy is performed in distinctive feature space, i.e. $ab:c:X$, with $a$, $b$, and $c$ given. For each language and each $n$, we show the best performing system in bold font. Scores which are statistically significantly better than scores for random sets of analogies are marked by an asterisk $^*$.  

5 Experiments

Correlation Our first experiment investigates the relationship between the geometries of embedding space and the distinctive feature space.

Let the embedding for phoneme $p$ be $\text{emb}(p)$, its distinctive feature vector $\text{feat}(p)$, and cosine similarity of vectors $u$ and $v$ be given by Equation 2.

\[
\text{sim}(u, v) = \frac{u^\top v}{|u| \cdot |v|} \quad (2)
\]

We measure the linear correlation of $\text{sim}(\text{emb}(p), \text{emb}(q))$ and $\text{sim}(\text{feat}(p), \text{feat}(q))$ over all unordered pairs of phonemes \{p, q\} (where $p \neq q$) using Pearson’s $r$. As a baseline, we compute the correlation of similarities of feature representations and random embeddings $\text{remb}(p)$. These are derived by randomly permuting the embeddings of phonemes. That is, $\text{remb}(p) = \text{emb}(q)$ for some random phoneme $q$.

Analogy Our second experiment investigates phoneme analogies. We first score four-tuples $(a, b, c, d)$ of phonemes using cosine similarity in embedding space as defined by Equation 3. This corresponds to a proportional analogy $a:b::c:d$.

\[
\text{score}(a, b, c, d) = \text{sim}(\text{emb}(b) - \text{emb}(a), \text{emb}(d) - \text{emb}(c)) \quad (3)
\]

We then evaluate the top 15, 30 and 100 four-tuples w.r.t. phonological analogy in distinctive feature space. Our evaluation is based on applying the transformation defined by the first two phonemes $a$ and $b$ on the third phoneme $c$ and measuring the Hamming distance of the result and the feature representation of $d$. For example, given tuple $(p, b, t, d)$, we get Hamming distance 0. This happens because $p$ is transformed to $b$ by changing the value of feature voice from 0 to 1. When the same transformation is applied to $t$, the result is $d$, which obviously has Hamming distance 0 with...
the fourth phoneme of the tuple in the embedding analogy. We restrict tuples in two ways: (1) all phonemes in the tuple have to be distinct symbols, and (2) all phonemes in the tuple have to be consonants or all of them have to be vowels.

As baseline, we randomly select 15, 30 or 100 phoneme tuples \((a, b, c, d)\). We then apply the transformation defined by \(a\) and \(b\) onto \(c\) and then compute the Hamming distance of the transformed image of \(c\) and the phoneme \(d\). We restrict these random tuples as explained above.

### 6 Results

Table 2 shows results for linear correlation measured by Pearson’s \(r\) for the similarity between phonetic representations and similarity of corresponding embedding vectors. Overall, the RNN encoder-decoder with embedding dimension 30 gives the best results. The correlation is the weakest for word2vec. However, all methods give a statistically significant positive correlation compared with random embeddings with p-value < 0.01 for appropriately chosen embedding dimension. For all three models: PPMI+SVD, word2vec and RNN encoder-decoder, there seems to be a tendency that higher dimension gives better correlation. This is not the case for PPMI+SVD for Spanish or word2vec for Turkish. However, in these cases, the results for all embedding dimensions are very similar. Figure 6 shows the correlation between cosine similarities of phoneme embeddings and the corresponding phonological feature representations.

Table 3 shows results for analogies as measured by average Hamming distance. Results are presented for the top 15, 30 and 100 analogies discovered by each of the systems. Overall, there is a strong trend that average Hamming distance increases in distinctive feature space when more (lower-ranked) analogies are considered in the embedding space. This is to be expected if the two spaces are coherent—as we include lower and lower ranked analogies and evaluate them, we expect them to be less fitting in the distinctive fea-
Figure 6: Scatter plots of cosine similarities of phonemes in feature space (x-axis) and embeddings space (y-axis). The figures present results for 30 dimensional PPMI+SVD, word2vec and RNN embeddings for Finnish, Turkish and Spanish, respectively. The red line represents the regression line.

The results both in comparing the geometry of the spaces learned and the alignment of analogies to distinctive features show a clear effect of distinctive features being aligned and discovered by distributional properties. The strength of the alignment appears to be somewhat language-dependent; in both Finnish and Turkish, vowel harmony effects are quite prominent and come out as many of the top-ranking analogies in an embedding space. In Spanish, by contrast, the correlation of the space is less robust, probably because there are fewer symmetrical phonological alternations witnessed in the data, although continuant alternation is a prominent one (/b/→/d/→/g/). Likewise, non-symmetric alternations in the data may distort the vector space to not align perfectly along distinctive feature lines. For example, while Finnish exhibits a /t/d/ alternation (katu/kadun; ‘street’ nominative/genitive) the corresponding analogical labial alternation in the embedding space is /p/v/ (apu/avun; ‘help’ nominative/genitive), not /p/b/, as one would assume by distinctive features. This is an interesting discovery since, while the analogy in the embedding space in this case does not correlate to the analogy in the feature space, this distortion of the embedding space of phonemes is arguably more “correct” than the feature-based expected one where /t:/→/p:b/. In fact, the /b/-phoneme is only present in loanwords in the Finnish data, and the spirantization seen in /p/v was historically present for the alveolar stop as well (/t/→/ð/). This analogy it-
self is an example of Finnish consonant gradation which manifests itself through many idiosyncratic alternations (Karlsson, 2008), some of which are clearly captured in symmetries in the embedding space. Hence, although such mappings are often present and prevents many analogies from being ‘perfect’ along distinctive feature lines, embedding spaces where such seemingly ‘incorrect’ analogies are drawn are in fact good representations for learning tasks such as morphological inflection, since they yield generalization power to task learning, i.e. learning of phonological alternations. This flexibility to learn a vector space representation that does not always strictly conform to distinctive features is then an advantage of the representations and partly explains their recent success (Cotterell et al., 2016, 2017) in learning inflectional patterns from examples.

8 Conclusion

We have presented a set of experiments on three languages that examine how distributional properties of phonetic segments contain information about regularities in the distinctive feature alternations present in the language. In particular, we have shown a significant correlation between embedding spaces learned from either co-occurrence and distinctive feature spaces. While such embeddings can be learned from raw data without any supervision, this correlation is consistently stronger if embeddings are learned and extracted from a recurrent neural network in conjunction with a supervised task of learning to inflect word forms. Apart from a holistic inspection of the embedding spaces, we also developed an experiment that measures how well phonological analogies can be performed using the embeddings learned. While the analogies do not perfectly correlate with similar analogies in distinctive feature space, it is clear that those distinctive features that play a part in prominent phonological alternations are also latently present in co-occurrence generalizations and can be seen in the learned embedding space.

References


Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for SIGMORPHON 2016 shared task on morphological reinflection. In SIGMORPHON.


