QUANTITATIVE METHODS FOR COMPARING FEATURAL REPRESENTATIONS

Ewan Dunbar, Gabriel Synnaeve, and Emmanuel Dupoux

Laboratoire de Sciences Cognitives et Psycholinguistique (ENS/EHESS/CNRS)
Ecole Normale Supérieure - PSL Research University
emd@umd.edu
Proceedings of International Congress of Phonetic Sciences (ICPhS 2015)

ABSTRACT

The basic representational hypothesis in phonology is that segments are coded using a universal set of discrete features. We propose a method for quantitatively measuring how well such features align with arbitrary segment representations. We assess articulatory, spectral, and phonotactic representations of English consonants. Our procedure constructs a concrete representation of a feature in terms of the pairs it distinguishes, and can be extended to any pair of representations to test the consistency of one with the individual dimensions of the other. We validate the method on our phonetic representations and then show that major natural classes are not well represented in the surface phonotactics.

Keywords: features; perception-production; quantitative methods; phonotactics

1. INTRODUCTION

Where do phonological features come from? Most feature sets attempt to make the classes that are "natural" to a learner featurally simpler; but it is also assumed that features are phonetically grounded. To assess how well natural classes are grounded in the phonetics, previous work has taken representations of segments based on articulation, acoustics, and tendency to group together crosslinguistically, and compared them qualitatively using visual inspection of a principal component analysis [7].

We propose a method for assessing this kind of alignment quantitatively. We limit ourselves to alignment with hand-defined natural classes, and demonstrate its use in finding what is coded in rich acoustic and articulatory representations. We then apply the method to the question of whether phonological features are emergent, i.e., substance-free and grounded only in language-specific phonological patterning [9, 6]. This would imply that phonological patterning should determine, and therefore align with, the phonological features. We find no evidence in favor of this hypothesis as it applies to English surface phonotactics aligning with standard place, manner, and voicing features on consonants.

2. DATA

To develop and test our method, we compared acoustic, articulatory, and phonotactically based representations of the consonant inventory of English. The acoustic representation of a phone was the corpus-wise average of 40 mel scaled filter banks, in a stack of 11 frames centered on the midpoint of each token of the category (25 ms windows spaced at 10 ms intervals), taken from TIMIT [3]. The articulatory representation of a segment was taken from the data reported on in [7]: vertical oral cavity distances estimated from ultrasound, plus vocal fold activity from EGG and oral and nasal airflow measurements, averaged over the productions of three trained phoneticians in V–V context for each of the three corner vowels. See [7] for details.

We used two different methods to extract a phonotactic representation for a segment without using features, both applied to the phonetic transcripts from the naturalistic Buckeye interview corpus [10]. The first uses a neural network (NN) approach where the surrounding phone sequence (plus and minus two segments) have to be predicted from a central phoneme [8]. The NN is a log–linear model composed of an input layer (of 39 phonemes) mapping to an embedding (of dimension 10) and then to the output layer (logistic regressions to 4×39 phonemes), and is trained with backpropagation with the logistic loss. The weights are used as a representation for the input phone. The second uses a singular value decomposition approach (SVD). We recorded relative frequencies of individual left-context phones and bigrams, right-context phones and bigrams, and of individual left-and-right contexts with a window of both one and two segments. We took the resulting table of relative frequencies and applied matrix factorization to reduce the dimension to 30. Both models ignored word boundaries. (An SVD model that included word boundaries was not meaningfully

Feature	Minimal phone pairs
Nasal	m–b, n–d, ŋ–g
Continuant	v–b, f–p, z–d, s–t
Coronal-Dorsal	d–g, t–k, n–ŋ
Coronal-Labial	d–b, t–p, s–f, z–v, n–m
Labial-Dorsal	b–g, p–k, m–ŋ
Voice	b–p, d–t, g–k, v–f, z–s, ʒ–ʃ

Table 1: Minimal pair sets for each of the six features.

different from these ones; we thus leave it out of further discussion.) Related approaches to pattern extraction from context are used in document processing [5]. Unlike the most widely used linguistic algorithm for learning phonotactics [4], neither presupposes any featural analysis.

Principal component analysis (PCA) was applied to rotate, center, and rescale all the feature representations, and reduction the dimension to the top 29 principal components (except for the NN features, which were kept at dimension 10). This is important, in order to correct for arbitrary differences in the scale of different dimensions (most obviously, between airflow and ultrasound measurements).

3. METHOD

Starting with a set of segments, we set up contrasts based on a fixed binary feature representation. We then evaluate various representations of the same set of segments to see how well these contrasts are captured. We focused on six featural contrasts for consonants: two manner features ([nasal], [continuant]), three place contrasts, ([coronal]–[dorsal], [coronal]–[labial], and [labial]–[dorsal]), and [voice]. Although for current purposes we set up these contrasts by hand, each one is meant to correspond to a feature contrast in a binary feature system.

3.1. Constructing representations of features

A feature is matched to a set of minimal pairs of phones that contrast in that feature. The sets corresponding to each of the features we selected are shown in Table 1.

These pairs are minimal in the sense that the two segments differ primarily on the given feature. We chose these sets as largely uncontroversial contrasts.

In fact, the notion of "primarily" or "contrastively" is underdetermined. Natural inventories are not symmetrical enough for there to be many sets which differ only in a single feature. For example, although [s] and [t] uncontroversially stand in the same relation as [f] and [p] in many languages, [f] is labiodental, while [p] is bilabial. This is often coded using the feature [distributed], a feature on which [s] and [t] do not differ.

Such additional differences between pairs can be deemed not to count towards minimal pairs. Ultimately, however, selection criteria are needed to determine which features count for which sets of segments. The contrastive hierarchy hypothesis [2] argues that the contrastive features for a segment should be determined on a language specific basis on the basis of the phonological patterning of that language. Alternatively, one could determine the relative importance of features to particular segments depending only on the other members of the inventory, an approach which has the advantage of requiring only an inventory table, and thus being easy to determine using objective statistical or information-theoretic criteria. The general problem is that of determining and justifying contrastive featural classes. Any attempt to pick out individual dimensions in one encoding for comparison with another must deal with this problem in one way or another, but we bracket the problem in order to demonstrate the core of our method.

Once a minimal phone pair set is constructed, each member is translated into a representation of the given feature contrast. Even though we constructed our minimal phone pairs by hand, we operate on the premise that they capture minimal specification differences in some particular feature representation. The full featural representations for the two phones are coded as vectors in this *classifying representation space*. Their vector subtraction becomes a concrete representation of that feature. We call each such subtraction vector a *frep*, for **rep**resentation of a **f**eature. We generated our vectors using a modified version of the system of [1] with most feature specifications irrelevant to the contrast being tested removed. (The exact feature system does not matter if the classes are chosen by hand, as we have done here, as our method simply uses the vectors to divide freps into those that represent the same versus different features; see below.)

After the freps are generated in the classifying representation space, equivalent freps are generated in the encoding being evaluated, the *test representation space*. The method is based on the idea that similar classifying representation freps should correspond to similar test representation freps. Alongside the freps for a feature, we take the phones in Table 1 and recombine them into all other possible minimal phone pairs (which can be seen as freps corresponding to other features). See Figure 1.

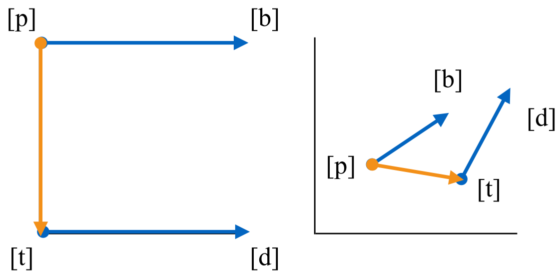

Figure 1: *Left (classification space):* two freps for [voice] (blue) and one frep for a different feature, in this case coding the coronal–labial distinction, (orange). *Right (test space):* the corresponding freps in a hypothetical encoding.

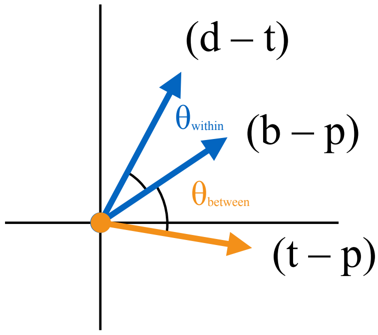

Figure 2: Our method compares the cosine of the angles between freps for the same feature and freps between different features, in the test representation space.

3.2. Comparing representations of features

For a given feature f, we compare freps for f with freps for f, and freps with non-freps for f, i.e., the other minimal phone pairs constructed out of the set of phones in Table 1, i.e., a set of relevant freps for other features. Freps for f are, by construction, more similar to each other in the classifying representation than they are to freps for any other feature g. Our test assesses whether this is also true in the test representation.

We use the absolute value of the cosine of the angle between two vectors as a frep similarity function in the test space:

$$\text{(1)} \quad \text{sim}(\vec{x}, \vec{y}) = \left| \frac{\vec{x} \cdot \vec{y}}{\|x\| \|y\|} \right|$$

Orthogonal freps will have similarity 0, while collinear freps will have similarity 1. In general, if f is well represented in the test representation, then we expect within-feature frep pairs (f frep versus f frep) to be more similar than between-vector frep pairs (f frep versus g frep). See Figure 2.

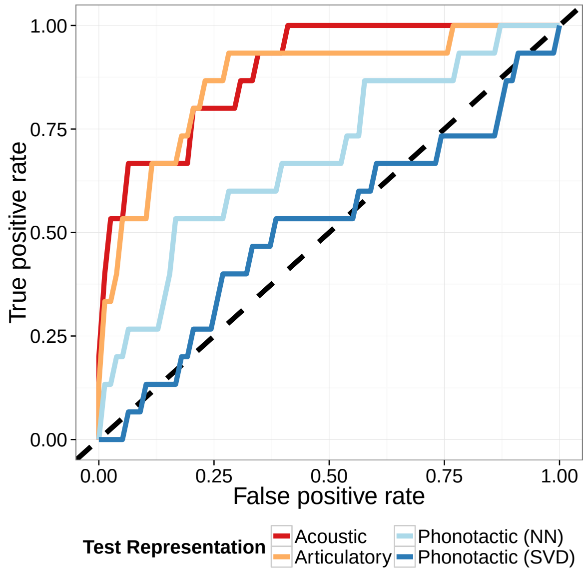

Figure 3: ROC curves for the four test representations for the feature [voice].

3.3. Alignment scores

All frep pair similarities in the test space are computed (both the f–f and the f–g pairs associated with a feature's minimal phone pair set). We assess how easily the similarity values are classified into within- versus between-feature pairs. We compute a receiver operating characteristic curve (ROC curve), and compute the *alignment* of the test representation with the given feature as the area under the curve (AUC). The ROC curve is given by the set of unique $\langle$false positive rate, true positive rate$\rangle$ pairs obtained from assessing every possible within/between classifier that is linear on similarity. An example for [voice] is shown in Figure 3.

The exact area under the curve is computed from these points. The maximum score is 1, and the minimum is 0; 0.5 indicates that the classifier is at chance. Values below 0.5 will occur if the test representation is better at capturing dimensions that are negatively correlated with the feature of interest. To the extent that the number of f–f reps is small compared to the number of f–g freps, f–f freps may also be distributed unevenly in the test space, which can give the appearance of negative (or positive) correlation with the feature by chance.

4. RESULTS

Alignment scores are shown in Figure 4. The acoustic representation performs very well for manner and voicing, but is poor for place features. The articula-

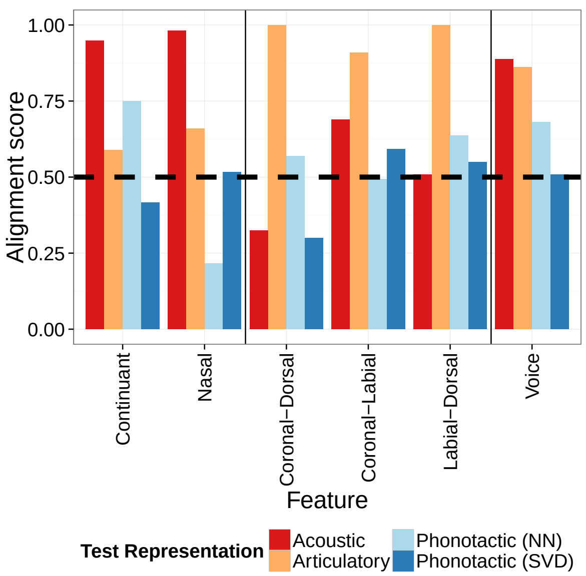

Figure 4: Alignment scores for all four test representations (0.5 is chance and 1.0 is the maximum); versus manner features ([continuant] and [nasal]), place contrasts ([coronal]–[dorsal], [coronal]–[labial], and [labial]–[dorsal]), and the feature [voice].

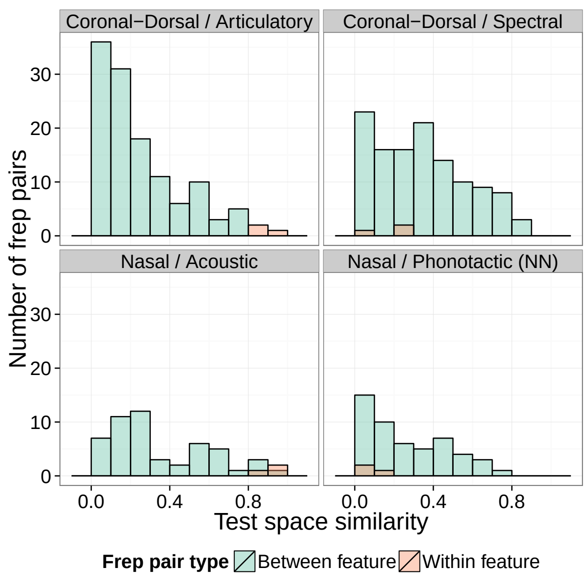

Figure 5: Within and between class frep pairs for good (upper and lower left) versus badly performing test representations (upper and lower right).

tory representation performs very well for place features and voicing, and poorly for manner features. The poor performance on manner may be due to the fact that this representation consists of duration normalized ultrasound and airflow data, which are not likely to capture these fine distinctions. Finally, neither representation of English phonotactics performs substantially above chance.

Some tests show actually below chance performance: for example, [coronal]–[dorsal] for the spectral representation, and [nasal] for one of the representations of phonotactics. Figure 5 compares the distribution of within and between feature similarities for these representations versus the acoustic representation, which performs well.

Although the number of pairs is small, the within-feature scores well are below the median, suggesting that the dominant dimensions in the test representations do indeed tend toward negative correlation with the features being tested.

5. CONCLUSION

We presented a new method for evaluating the grounding of phonological representation in acoustic, articulatory and phonotactic space. The results on acoustic and articulatory are consistant with expectations about the grounding of place and manner features. The fact that the phonotactics of English align so poorly with basic natural classes prima facie challenges emergent feature theories. We have certainly not exhausted the space of cognitively reasonable phonological pattern based representations; nevertheless, given that at least some of the features we tested are phonologically active in English (for example, place, in nasal place assimilation, and voicing, in coda obstruent voicing restrictions), it is worrying that we do not see them robustly attested.

It is unsurprising that at least the core of the standard phonological feature system for consonants is consistent with major dimensions of phonetic variability, since standard phonological features are at least partly phonetically interpretable by design. The general procedure we have presented does not allow us to evaluate the hypothesis that any of these features is cognitively active; rather, it allows us to assess which sources of information are available in acquisition and processing, if any, that would support their being cognitively active. That we have a quantitative measure of this is an advance over previous research.

Future work should explore general procedures for working with a classifying representation to automatically determine the dimensions of greatest contrast, keeping in mind that this may be different depending on a segment's location in the classifying space. With such a tool available, it would become straightforward to compare arbitrary pairs of representations. An extension to continuous classifying representations can be made, for example, to summarize how clearly important articulatory parameters are easily recoverable from the spectrum without needing to narrowly define what the spectral signature should look like. This could be done by discretizing and using the procedure described here, or by evaluating correlations rather than two-way classification performance.

6. REFERENCES

[1] Clements, G. N., Halle, M. 1983. Problem book in phonology.

[2] Dresher, B. E. 2009. *The contrastive hierarchy in phonology* volume 121. Cambridge University Press.

[3] Garofolo, J. S., Consortium, L. D., others, 1993. *TIMIT: acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium.

[4] Hayes, B., Wilson, C. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry* 39(3), 379–440.

[5] Landauer, T. K., Dumais, S. T. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2), 211.

[6] Mielke, J. 2008. *The emergence of distinctive features*. Oxford University Press.

[7] Mielke, J. 2012. A phonetically based metric of sound similarity. *Lingua* 122(2), 145–163.

[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 3111–3119.

[9] Morén, B. 2006. Consonant–vowel interactions in serbian: Features, representations and constraint interactions. *Lingua* 116(8), 1198–1244.

[10] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., Raymond, W. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1), 89–95.