PHACTS about activation-based word similarity effects

Basilio Calderone
CLLE-ERSS (UMR 5263) CNRS &
Université de Toulouse-Le Mirail
31058 Toulouse Cedex 9, France
basilio.calderone@univ-tlse2.fr

Chiara Celata
Scuola Normale Superiore
Laboratorio di Linguistica
56126 Pisa, Italy
c.celata@sns.it

Abstract

English phonotactic learning is modeled by means of the PHACTS algorithm, a topological neuronal receptive field implementing a phonotactic activation function aimed at capturing both local (i.e., phonemic) and global (i.e., word-level) similarities among strings. Limits and merits of the model are presented.

1 Introduction

Categorical rules and probabilistic constraints of phonotactic grammar affect speakers' intuitions about the acceptability of word-level units in a number of experimental tasks, including continuous speech segmentation and word similarity judgment. Several sources of information contribute to phonotactic generalization, including sub-segmental properties, segment transition probabilities, lexical neighborhood effects; all these factors have been independently or jointly modeled in several recent accounts of phonotactics and phonotactic learning (Coady and Aslin, 2004; Vitevitch, 2003; Vitevitch and Luce, 2005; Hayes and Wilson, 2008; Albright, 2009; Coetzee, 2009).

In this study, we explore the word level phonotactics in terms of a function of 'phonotactic activation' within a PHACTS environment (Celata et al., 2011). PHACTS is a topological neuronal receptive field implementing an n-gram sampling estimate of the frequency distribution of phonemes and a sub- lexical chunking of recurrent sequences of phonemes. Once this phonotactic knowledge has been developed, the model generalizes it to novel stimuli to derive activation-based representations of full lexical forms, thus mirroring the contribution of lexical neighborhood effects. Then the similarity values for pairs of words and non-words can be calculated.

2 PHACTS: the model

PHACTS (for PHonotactic ACTivation System) is based on the principles of a Self-Organizing Map (SOM) (Kohonen, 2000), an associative memory algorithm which realizes low-dimensional (generally, bi-dimensional) representations of a multidimensional input space.
PHACTS simulates the formation of phonotactic knowledge in the mind of a speaker, who is exposed to a stream of phonological words and gradually develops a mental representation of the statistical regularities shaping the phonotactics of a given language. The model also performs lexical generalizations on the basis of the phonotactic knowledge developed in the training phase.

The physical structure of PHACTS is defined by a set S (with finite cardinality) of neurons n_{jk} with $1 \leq j \leq J$ and $1 \leq k \leq K$ arranged in a bi-dimensional grid of $S = \{n_{11}, n_{12}, \ldots n\}$, $\|S\| = JK$. Each neuron in the grid corresponds to a vector (the so-called prototype vector) whose dimension is equal to the dimension of the input data vector. At the beginning of the learning process, the prototype vectors assume random values while, as learning progresses, they change their values to fit the input data.

PHACTS works according to the two following phases: i) the training phase, where language-specific phonotactic knowledge is acquired; ii) the lexical generalization phase.

Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss, pages 33–37,
Avignon, France, April 24 2012. ©2012 Association for Computational Linguistics

2.1 Training phase: the acquisition of phonotactic knowledge

At the beginning, each input word iteratively hits the system. For any iteration, the algorithm searches for the *best matching unit* (BMU), that is, the neuron which is topologically the closest to the input vector i and which is a good candidate to represent the input data through the prototype vector. The search for the BMU is given by maximizing the dot product of i and u_{jk} in the *t*-th step of the iteration:

$$BMU((i)t) = \arg\max_{jk}(\mathbf{i}(t) \cdot \mathbf{u}_{jk}) \qquad (1)$$

In other terms, the $BMU((i)t)$ is the best aligned prototype vector with respect to the input i. After the BMU is selected for each i at time t, PHACTS adapts the prototype vector u_{jk} to the current input according to the topological adaptation equation given in (2):

$$\Delta u_{jk}(t) = \alpha(t)\delta(t)[i(t) - u_{jk}(t-1)] \qquad (2)$$

where $\alpha(t)$ is a *learning rate* and $\delta(t)$ is the so-called *neighborhood function*. The *neighborhood function* is a function of time and distance between the BMU and each of its neighbors on the bi-dimensional map. It defines a set of neurons around the that would receive training, while neurons outside this set would not be changed. In our model the *neighborhood function* is defined as a Gaussian function.

The α parameter controls for the elasticity of the network, and δ roughly controls for the area around each best matching where the neurons are modified. The initial value of both parameters is set heuristically and in general decreases as long as the learning progresses. In order to facilitate a training convergence, we set $\alpha \rightarrow 0$ and $\delta \rightarrow 0$ as $t \rightarrow 0$. PHACTS performs a vector mapping of the data space in input to the output space defined by the prototype vectors u_{jk} on the bi-dimensional grid of neurons S.

2.1.1 The data: Type and token frequency in PHACTS

For the present simulations, PHACTS was trained on a portion of the CELEX English database (Baayen et al., 1995), and specifically on 8266 English word types phonologically transcribed and provided with their frequency of occurrence (only the words with token frequency > 100 were selected). Each phoneme was phonologically encoded according to a binary vector specifying place, manner of articulation and voicing for consonants, roundedness, height and anteriority for vowels. The bi-dimensional map was 25 X 35 neurons, and thus $S = 875$. Input words were sampled according to i for PHACTS is constituted by the input training words with a n-gram sampling window (with *n* spanning up the length of the longest word).

During the training phase, the map takes into account the global distribution of the n-grams in order to realize the topological activations of the phonotactic patterns ('phonotactic activation'). Both token frequency (i.e., the number of occurrences of specific n-grams) and type frequency (i.e., the number of all members of an n-gram type as defined by phonological features shared; for instance, */tan/* and */dim/* are two realizations of the trigram type *stop+vowel+nasal*) play a key role in phonotactic activation. By virtue of being repeatedly inputted to the map, a high token frequency n-gram will exhibit high activation state in the map. Low token frequency n-grams, however, will exhibit activation on the SOM only if they share phonological material (namely, phonemes or features) with high token frequency n-grams. Type frequency generates entrenchment effects in the map; high type frequency n-grams will occupy adjacent positions on the bi-dimensional map, thus defining clear phonotactic clusters. For these reasons, PHACTS differ sharply from current models of phonotactic learning, where only type frequencies are assumed to play a role in phonotactic generalization (and formalized accordingly). (Albright, 2009)

2.2 N-gram generalization and lexical generalizations

Once PHACTS has been exposed to an input of phonologically-encoded n-grams , an activation-based representation of unseen words can be derived. This phase implements a linear thresholded function d in which each neuron ŤfiresŢ as a function of its activation with respect to the (unseen) n-grams. In this sense each neuron acts as a 'transfer function'Ţ of an activation weight depending on the alignment between the unseen n-gram vector and the best aligned n-gram prototype vector.

Lexical generalization in PHACTS is therefore a word-level transfer process whereby the activation values of each word n-gram are summed according to equation [4]:

$$F_{\text{PHACTS}}(x) = \sum_{jk} \Phi(x) \tag{3}$$

The cumulative action of n-gram activations realizes a distributed representation of the word in which both phonological similarity (at the string level), and token frequency effects for phonotactic patterns are taken into account.

Being based on an associative memory learning of phonological words inputted by a n-gram sampling window, PHACTS develops topological cumulative memory traces of the learned words in which phonotactic activations emerge as the results of repeated mnemonic superimpositions of n-grams. This aspect is crucial for a distributional analysis of the morphotactic salience in a given language. In this direction, PHACTS was successfully implemented in the modeling of the micro- and macro-phonotactics in Italian (Calderone and Celata, 2010). By micro-phonotactics we mean sequential information among segments (e.g., the fact that, in the specific language, a phonological sequence, such as */ato/*, differs from similar sequences, such as */uto/*, */rto/*, and */atu/*). By macro-phonotactics we mean positional information within the word, i.e., sub-lexical (or chunk) effects (e.g., the fact that word-initial */#ato/* is different from word-medial */-ato-/*, as well as from word-final */ato#/*). In English language as well, PHACTS seems to distributionally distinguish a positional relevance for highly attested phonological sequences such as */ing/*. Figure 1 reports the phonotactic activation states outputted for the sequence */ing/* in initial and final word position (training corpus and parameters described in 2.1.1).

3 The experiments

According to the literature, the speakers in judging the wordlikeness of isolated non-words rely mainly on a grammar-based phonotactic knowledge and enhance the correspondence among types of strings (e.g., segmental features and onset and coda constituency). In doing so, they establish connections between each non-word and the

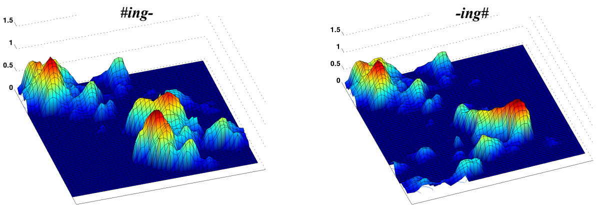

Figure 1: Phonotactic activation states for the sequence *#ing-* (initial word position) and *-ing#* (final word position)

neighborhood of all attested and unattested (but phonotactically legal, i.e., potentially attested) strings of their language. This must be a computationally hard task to accomplish even when no time restrictions are imposed, as in traditional wordlikeness experiments (since (Scholes, 1966) onward). In this experiment, we want to verify whether such task can be modeled in PHACTS and whether the vector representation of words outputted by PHACTS may represent a solid basis for this type of phonotactic evaluation. To evaluate PHACTS's ability to reproduce the typicality patterns produced by the speakers in judging the 'Englishness' of isolated strings, we had to derive a similarity value among each string and some counterpart in the English lexicon, as explained with more details below. We used 150 non-words, which were randomly selected from the list of 272 non-words of Bailey and Hahn (2001, B & H henceforth).

In that study, pronounceable non-words were created, either 4- or 5-phoneme long, differing from their nearest real word neighbor by either one or two phonemes (in terms of substitution, addition or subtraction). In the former case they were called near misses, in the latter case they were called isolates. 22 isolates and 250 near misses around the isolates were used in the B & H's study; 24 English speakers were asked to judge the 'Englishness' of the non-words that were individually presented in their orthographic and auditory form. The 150 non-words used in the present experiment were selected from among the near misses only. PHACTS was asked to derive the cosine value between the vector representations of each non- word and the corresponding real English words composing its neighbor family (according to the lists provided in B & H). The total number of string pairs was 1650 (the average number of neighbors for each non-word being 11). Then, an average cosine value was calculated for each of the 150 non-words. The average cosine value was assumed to reflect the phonotactic acceptability of each non-word with respect to their real word neighbors and therefore, to approximate the speakers' typicality judgment of isolated non-words. An edit distance calculation (normalized by the length of the two strings) was performed for the same 1650 pairs of non-words. Since the neighbors were all selected by adding, subtracting or modifying one phoneme from their reference non-words, the edit distance values were expected not to vary to a large extent. In the edit distance algorithm, values range from 0 to 1 according to the degree of the similarity between the two strings As expected, the distribution of the edit distance values was not uniform and the 1650 string pairs elicited a very small range of edit distance values. In total, 96% of cases elicited only four different edit distance values (namely, 0.83, 0.87, 0.93 and 0.97); the remaining 4% elicited three different values which were all higher than 0.7.

The cosine values outputted by PHACTS for the same string pairs were evaluated with respect to the calculated edit distances. As in the case of the edit distance algorithm, cosine values close to 1 indicate high similarity while values close to 0 indicate low similarity. As in the case of the edit distances, the cosine values were asymmetrically distributed, highly skewed to the right (for high similarity values). The global range of the distribution of values was similar for the two algorithms (spanning from 0.7 to 0.99). However, compared to the sharpness of the edit distance results (see Figure 2), PHACTS's output included subtler variations across comparisons, with fine distinctions distributed over a continuous range of values. The edit distance and the cosine values turned out to be correlated with $r = 0.465$. Although the nature of the difference between PHACTS's output and the edit distance algorithm should be better evaluated with respect to a more varied data set, also including pairs of very dissimilar strings, we could preliminarily conclude that the cosine value calculated by PHACTS for pairs of activation-based string representations did not correspond to an edit distance calculation.

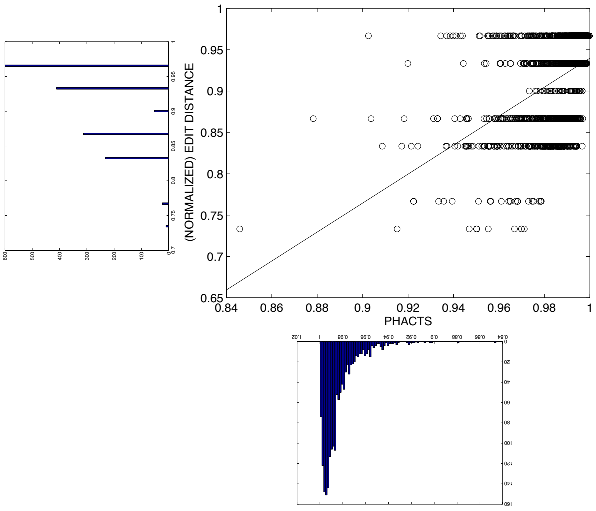

Figure 2: Correlation scatterplot and distribution histograms of the edit distance and PHACTS values for the B & H's materials

We further verified whether PHACTS cosine values could approximate the perceived phonotactic distance between two strings, as it is calculated by the speaker when (s)he is asked to judge the phonotactic acceptability of an isolated non-word. To test this hypothesis, the average cosine value of each non-word was correlated with the corresponding acceptability rating produced by the English subjects in the B & H's work. The Spearman's rank correlation between speakers' ratings and the (exp-transformed) cosine values was $\rho = .216, p < .01$. Although statistically significant, the correlation coefficient was rather low and revealed that the observed and simulated behaviors overlapped only to a limited extent. In particular, PHACTS did not reach a span of phonotactic acceptability as large as the speakers appeared to produce (with ratings comprised between 2.1 and 6.5).

In conclusion, PHACTS-based word similarity calculation appeared not to produce a reliable ranking of strings according to their phonotactic wellformedness. On the other hand, it did produce a fine-grained distributed representation of word in which both phonological similarity and token frequency effects for full forms seemed to define phonotactic activations of highly attested phonological sequences. This kind of representation differed from raw calculations of the number of operations required to transform a string into another.

Experimental protocols for modeling word similarity in PHACTS are currently under investigation.

References

Adam Albright. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1):9–41.

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The celex lexical database. release 2 (cd-rom). *Philadelphia: Linguistic Data Consortium, University of Philadelphia: Linguistic Data Consortium, University of Pennsylvania*.

Basilio Calderone and Chiara Celata. 2010. The morphological impact of micro- and macro-phonotactics. computational and behavioral analysis (talk given). In *14th International Morphology Meeting*, Budapest, 13-16 May.

Chiara Celata, Basilio Calderone, and Fabio Montermini. 2011. Enriched sublexical representations to access morphological structures. a psycho-computational account. *TAL-Traitement Automatique du Langage*, 2(52):123–149.

Jeffry A. Coady and Richard N. Aslin. 2004. Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, 89:183–213.

Andries W. Coetzee. 2009. Grammar is both categorical and gradient. In S. Parker, editor, *Phonological Argumentation: Essays on Evidence and Motivation*. Equinox.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Teuvo Kohonen. 2000. *Self-Organizing Maps*. Springer, Heidelberg.

Robert J. Scholes. 1966. *Phonotactic Grammaticality*. Mouton.

Michael S. Vitevitch and Paul A. Luce. 2005. Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, 52(2):193–204.

Michael S. Vitevitch. 2003. The influence of sublexical and lexical representations on the processing of spoken words in english. *Clinical Linguistics & Phonetics*, 17:487–499.