Local and non-local dependency learning and emergence of rule-like representations in speech data by Deep Convolutional Generative Adversarial Networks

Gašper Beguš[a]

[a] *Department of Linguistics, University of California, Berkeley, 1203 Dwinelle Hall #2650, Berkeley, CA 94720*

Abstract

This paper argues that training Generative Adversarial Networks (GANs) on local and non-local dependencies in speech data offers insights into how deep neural networks discretize continuous data and how symbolic-like rule-based morphophonological processes emerge in a deep convolutional architecture. Acquisition of speech has recently been modeled as a dependency between latent space and data generated by GANs in Beguš (2020c), who models learning of a simple local allophonic distribution. We extend this approach to test learning of local and non-local phonological processes that include approximations of morphological processes. We further parallel outputs of the model to results of a behavioral experiment where human subjects are trained on the data used for training the GAN network. Four main conclusions emerge: (i) the networks provide useful information for computational models of language acquisition even if trained on a comparatively small dataset of an artificial grammar learning experiment; (ii) local processes are easier to learn than non-local processes, which matches both behavioral data in human subjects and typology in the world's languages. This paper also proposes (iii) how we can actively observe the network's progress in learning and explore the effect of training steps on learning representations by keeping latent space constant across different training steps. Finally, this paper shows that (iv) the network learns to encode the presence of a prefix with a single latent variable; by interpolating this variable, we can actively observe the operation of a non-local phonological process. The proposed technique for retrieving learning representations has general implications for our understanding of how GANs discretize continuous speech data and suggests that rule-like generalizations in the training data are represented as an interaction between variables in the network's latent space.

Keywords: neural networks, behavioral experiments, machine learning, learning biases, speech, morphology

1. Introduction

The discussion between connectionist and symbolic approaches to language and human cognition in general has long been in the focus of computational cognitive science (Rumelhart et al. 1986; McClelland et al. 1986; Marcus 2001, i.a.). Phonetic and phonological data are uniquely appropriate for addressing this problem. Over a century-long tradition of scientific study of acoustic and perceptual phonetics (for an overview, see MacMahon 2013) that deals with physical properties of speech sounds provides a solid understanding of the continuous data that hearing infants acquire language from: raw acoustic speech. Phonology is the study of how humans analyze, discretize,

Email address: begus@berkeley.edu (Gašper Beguš)

self-organize, and manipulate continuous speech data into discretized mental representations called *phonemes.* The scientific study of phonology, too, has an over-a-century long history (for an overview, see van der Hulst 2013), which resulted in a solid understanding of local and non-local discrete dependencies in human speech. Phonetic and phonological data and analysis are thus uniquely appropriate for probing what deep convolutional networks can and cannot learn, how discrete representations can emerge in deep neural networks, and how their performance can be paralleled to human behavior. Despite these advantages, the majority of neural network interpretability studies focus on non-linguistic visual data or syntactic/semantic levels, the latter of which lack a continuous component.

Computational models of speech acquisition have a long history. The majority of models, however, operate with abstract and already discretized data rather than raw acoustic inputs (McClelland and Elman, 1986; Gaskell et al., 1995; Plaut and Kello, 1999). Deep neural network models of phonetic and phonological data operating with raw acoustic inputs emerged only recently. Several proposals model phonetic learning with deep autoencoder models (Räsänen et al., 2016; Alishahi et al., 2017; Eloff et al., 2019; Shain and Elsner, 2019). Autoencoders learn to reduce data and encode data distributions in latent representations: they are trained on reproducing inputs by generating outputs from a reduced latent space. Inputs are thus directly connected to the outputs with an intermediate latent space that is reduced in dimensionality. Clustering analyses on the latent space show that the networks trained on phonetic data learn approximations of phonetic features from phonetic similarity (Räsänen et al., 2016; Alishahi et al., 2017; Eloff et al., 2019; Shain and Elsner, 2019).

While the reduced dimensionality in the autoencoder architecture approximates phonetic features and phonetic similarity, the proposals do not model phonological processes. The human language learner has to acquire not only the identity of individual sounds based on acoustic similarity (as approximately modeled by the proposals using the autoencoder architecture), but also to manipulate those sounds in a given phonetic context. For example, a voiceless bilabial stop /p/ in English can surface as aspirated [p^h] (produced with aspiration or a puff of air) before stressed vowels or as unaspirated [p] (without aspiration) if a fricative [s] precedes it. A minimal pair illustrating this distribution is [ˈ**p$^{\mathbf{h}}$**ɪt] ‘pit’ and [ˈs**p**ɪt] ‘spit’. The learner needs to learn not only to output voiceless bilabial stop, but also to shorten the aspiration time (VOT) when an [s] precedes it. Autoencoders are also trained on replicating output data as closely as possible to the input data, which is not desirable in models of language acquisition. While dimensionality reduction in autoencoders is unsupervised, input-output pairing is not.

To model phonetic learning simultaneously with the learning of simple allophonic processes, Beguš (2020c) proposes that speech acquisition can be modeled as a dependency between the latent space and generated data in the Generative Adversarial Networks. Generative Adversarial Networks (GAN), first proposed by Goodfellow et al. (2014), have not been used for modeling language acquisition, despite several advantages that this architecture features for computational models of language learning. GAN models are unsupervised and fully generative, which means that a deep convolutional network outputs innovative data that have no direct link to the training data (unlike, for example, in the autoencoder architecture). In other words, deep convolutional networks in the GAN architecture need to learn to output data from some random distribution.

Beguš (2020c) argues that deep convolutional networks in the GAN architecture encode discretized phonetic and phonological representations in the latent space. A computational experiment is conducted on a GAN implementation for audio (as proposed in Donahue et al. 2019 based on Radford et al. 2015) by training the networks on an phonologically local allophonic distribution in English, where voiceless stops surface as aspirated word-initially before a stressed vowel (e.g. in [ˈp^{h}ɪt] ‘pit’), except if a sibilant [s] precedes the stop (e.g. in [ˈspɪt] ‘spit’). The network learns the allophonic distribution and encodes phonetically and phonologically meaningful features in its latent space.

Based on this local allophonic distribution, Beguš (2020c) proposes a technique for identifying and manipulating variables in the latent space in the GAN architecture that correspond to desired phonetic and phonological representations. Beguš (2020c) argues that the network uses a subset of latent variables to encode presence of a sound in the output (e.g. [s]). By manipulating the identified variables, especially well beyond the training range (as proposed in Beguš 2020c), we can actively force the sound in and out of the generated outputs. Moreover, a linear interpolation of the chosen latent variables from marginal values results in almost linear reduction of the amplitude of the frication noise of [s] — a linguistically meaningful unit (Beguš, 2020c).

The goal of this paper is to argue that using the technique proposed in Beguš (2020c), we can model not only simple allophonic processes, such as English deaspiration, but also local and non-local phonological processes that are based on what would be approximated as morphology (morphophonological alternations) that resemble rule-like behavior. We also argue that we can parallel human behavioral experiments with performance of the deep convolutional networks that are trained on the same data as used in behavioral experiments. In general, natural languages strongly prefer local over non-local processes, both in phonology and on other levels such as morphology and syntax (Finley, 2011, 2012; McMullin and Hansson, 2019; White et al., 2018). In fact, the vast majority of phonological processes in the world's languages are local (Finley, 2011), with only a few processes, such as harmony, operating on non-adjacent sounds. Behavioral experiments show that local processes are easier to learn than non-local processes (Finley, 2011, 2012; McMullin and Hansson, 2019; White et al., 2018). In this paper, we test the learning of local and non-local phonological dependencies, and show that local processes (such as postnasal or intervocalic devoicing) are easier to learn for the networks than non-local vowel harmony. We parallel success rates in the computational model to behavioral data — an artificial grammar learning experiment in which human subjects are trained on the same data (Section 4). This type of combining artificial grammar learning experiments and computational models has the potential to reveal similarities in learning biases between human subjects and deep convolutional networks, and shed light on how domain-general learning biases that require no language-specific mechanisms can result in the typological prevalence of local processes and the rarity of non-local processes.

Specifically, we test the learning of non-local vowel harmony and several devoicing patterns. Vowel harmony is a phonological process, usually non-local, in which a vowel becomes more similar to another vowel in a word. For example, the plural morpheme in Turkish surfaces as [lɑr] after root vowels that are back and as [lɛr] if the root vowel is front Kabak (2011): [d**ɑ**l-l**ɑ**r] 'places' and [j**e**r-l**e**r] 'branches' (Kabak, 2011).

In formal phonological analysis, phonological computation is formalized with rewrite rules that operate as symbolic feature manipulation (Chomsky and Halle, 1968). As argued by Marcus et al. (1999) and several other works (Chomsky and Halle 1968; Heinz 2010; Berent 2013, i.a.), "algebraic rules" are required to derive a set of surface outputs such as Turkish [d**ɑ**l-l**ɑ**r] and [j**e**r-l**e**r] from stored inputs. The stored mental representation of the prefix can be posited as /lar/. The role of phonological grammar is to derive the two surface forms (outputs) from the stored mental representation (input).

Sounds are represented with matrices of binary features that distinguish meaning (e.g. [+syllabic, + front] means a front vowel). Vowel harmony can be formalized with a simple rewrite rule (in 1) that identifies vowels ([+syllabic]) and assigns the same value (α) of feature [±front] as in the vowel that follows it (interrupted by any number of consonants C_0). The formalism is illustrated in (1).

$$[+\text{syllabic}] \rightarrow [\alpha\ \text{front}]/\underline{\qquad}\text{C}_0[\alpha\ \text{front}] \tag{1}$$

The discussion of symbolic representation vs. connectionism has a long tradition in phonology. An influential proposal called Optimality Theory models phonology as an input-output pairing rather than a rule-based symbolic representation (Prince and Smolensky, 1993/2004; Legendre et al., 1990). Optimality Theory was directly influenced by earlier work on connectionism. Vowel harmony within this framework is modeled with the Agreement-by-correspondence proposal (Hansson, 2010; Rose and Walker, 2004): two sounds (such as the two vowels [ɑ] in Turkish [d**ɑ**l-l**ɑ**r]) are in correspondence and share features, which, through surface optimization in the grammar, results in a harmonious process. Several independent facts support the approach of input-output optimization in phonology. However, both Optimality Theory and other proposals in phonology using neural networks (McClelland and Elman, 1986; Gaskell et al., 1995; Plaut and Kello, 1999) model local and non-local phonology with pre-assumed levels of abstraction, meaning that learning is not modeled from raw acoustic data but is already pre-discretized or requires language-specific mechanisms.

We argue that approximates to rule-based behavior emerge in deep convolutional networks even without any pre-assumed levels of abstraction (the networks are trained on raw acoustic inputs) and when models contain no language-specific parameters. The network discretizes the representation of a prefix in the output and uses only one latent variable (out of 100) to encode the presence of the prefix. Equivalents to non-local phonological rules emerge from an interaction between the variable that represents the prefix and a variable that generates some desired phonological process. We also argue that the same data used for training in the GAN architecture can be used to test phonological learning in artificial grammar learning experiments in human subjects. In fact, the paper argues that training GANs on relatively few data points yields, somewhat surprisingly, highly informative results (Section 3.1). This observation should open numerous opportunities for paralleling performance in deep neural networks and behavioral outcomes of artificial grammar learning experiments with human subjects. Finally, we outline a procedure to observe how the network learns dependencies as the training progresses and claim that the generator's search through the space of phone-level combinations are linguistically interpretable (Section 3.2).

2. Materials

2.1. Model

The main characteristic of Generative Adversarial Network architecture (Goodfellow et al., 2014), and more specifically the DCGAN proposal by Radford et al. (2015), are two deep convolutional neural networks that are trained in a minimax setting. The Discriminator learns to estimate realness of the data and minimize its own error rate. The Generator network learns to output data from a set of latent variables and maximize the Discriminator network's error. Initially, the Generator network produces noise, but as training progresses it becomes increasingly more successful in outputting data such that the Discriminator becomes less successful in distinguishing actual from generated data.

The majority of GANs are trained on two-dimensional visual data; a shift to apply the architecture to the audio domain has occurred only recently with the work of Donahue et al. (2019) (WaveGAN). The model in Donahue et al. (2019), used for training here, is based on the DCGAN architecture (Radford et al., 2015) and features most of the same hyperparameters. The two main differences are that the Generator involves an additional layer and takes a one-dimensional input that corresponds to approximately 1 second of audio. The cost function is taken from the Wasserstein GAN proposal with gradient penalty (WGAN-GP) (as proposed in Arjovsky et al. 2017 and Gulrajani et al. 2017). For all specifications of the model, see Donahue et al. (2019).

Beguš (2020c) proposes a technique for exploring learning representations in deep convolutional networks. First, variables that correspond to meaningful phonetic and phonological representations are identified. For example, the network is trained on #$\mathrm{T^hV}$ and #sTV sequences from TIMIT

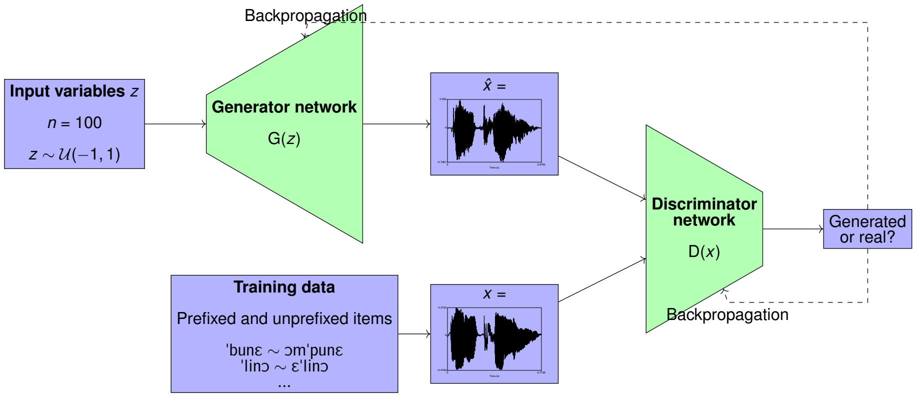


Figure 1: The GAN architecture schematized from (Goodfellow et al., 2014; Radford et al., 2015; Donahue et al., 2019) used in this paper with training data as described in Section 2.2.

(e.g. [p^{h}æ] and [spæ]) and learns the conditional distribution: it mostly outputs short VOT (no aspiration) if an [s] precedes the stop and long VOT (aspiration) if no [s] precedes it. However, the Generator's outputs are not simply replications of its input: in about 12% of outputs, the stop after an [s] is aspirated ([sp^{h}æ]) and the VOT duration is longer than in any #sTV sequence in the training data. Additionally, the network occasionally outputs innovative sequences that lack a stop (e.g. #sV) or concatenate two stops (e.g. #TTV). In other words, the Generator learns the conditional allophonic distribution, but imperfectly so (Beguš, 2020c). The outputs with long VOT (aspiration) in the [s]-condition parallel stages in language acquisition: language-acquiring children also occasionally output stops with long VOT (aspiration) in the [s]-condition (Bond and Wilson, 1980).

In addition to observing learning in the GAN architecture with surface forms, we can identify individual latent variables that correspond to phonetic and phonological representations. Beguš (2020c) proposes a technique for identification of the variables by regressing the annotated outputs to the randomly sampled latent space. Predictions of several regression models are tested in Beguš (2020c) to avoid assumptions of linearity: generalized additive models with various shrinkage techniques, linear logistic regression, Lasso logistic regression, and random forest models. The technique identifies latent variables (z; see Figure 1) that correspond to presence of [s] in the output. Moreover, it is shown that the relationship between the individual latent variables (e.g. those identified as representing [s]) and the generated data are often linear, even when non-linear regression is used for testing.

Given this linear relationship, we can identify variables that correspond to a desired phonetic property and identify whether the property correlates with positive or negative values of the variable. Individual z-variables are uniformly distributed during the training with the interval $(-1, 1)$. When set to a value identified as corresponding to presence of a desired phonetic feature, the output contains a significantly higher proportion of this property. Crucially, Beguš (2020c) shows that manipulating the identified variables beyond the values in the training range $(-1, 1)$, such as to ± 4.5, results in an increased amplitude of the desired phonetic representation. In other words, as we interpolate a variable identified as representing an [s] in the output, the amplitude of [s] increases or decreases. We can thus actively force a phonetic or phonological feature in the output. That the proposed technique indeed identifies variables corresponding to the presence of [s] is suggested by an independent generative test in Beguš (2020c). While explorations of latent space and representation learning in GANs have been conducted before on visual data (Radford et al., 2015), the proposals, to the author's knowledge, do not use single variables to explore their meaningful equivalents and do not utilize interpolation to extreme values beyond the training range.[1]

Beguš (2020c) thus argues that the Generator network learns a local allophonic distribution as well as learns to encode phonetic and phonological representations with a subset of variables in the latent space. While the Generator network represents [s] in the latent space with a subset of variables in Beguš (2020c), the cutoff between variables associated with presence of [s] and the rest of the latent space is not completely categorical. The Generator network does not associate the presence of [s] with a single variable: seven z-variables are associated with the representation of [s]. There is a notable cutoff between the regression estimates of the seven highest variables and the rest of the latent space, but the difference is not substantial or categorical. Training data in Beguš (2020c) is sliced from TIMIT (Garofolo et al., 1993), which is considerably more variable than the training data in this experiment. As is argued in Section 4, discretization of some phonological representation (e.g. presence of the prefix) is substantial in the current experiment. It appears that less variable data results in a more rapid discretization.

2.2. Data

The training data contain evidence for one non-local phonological process — vowel harmony — and four local processes: (i) post-nasal devoicing ([ˈ**b**ɑlu] ∼ [ɔmˈ**pʰ**ɑlu]), (ii) post-nasal occlusion with devoicing of voiced stops ([ˈ**v**iɹə] ∼ [ɛmˈ**pʰ**iɹə]), (iii) intervocalic devoicing ([ˈ**b**ulɔ] ∼ [ɔˈ**p**ʰulɔ]), and (iv) intervocalic fricativization with devoicing ([**b**ɔɹə] ∼ [ɔˈ**f**ɔɹə]). These processes are triggered by prefixes; the training data thus contain bare (unprefixed) and prefixed forms of lexical items of the shape (PREFIX-)CVCV and (PREFIX-)CVC (C = consonant, V = vowel), e.g. [ˈɹinu] ∼ [ɛnˈɹinu]. The items are all nonce words in English, so that the same dataset can be used in the behavioral experiment with human subjects (Section 4).

2.2.1. Non-local processes

Non-local vowel harmony is triggered by the first vowel of the base (unprefixed) form and results in two different vowel qualities of the prefix, [ɛ] and [ɔ]. The descriptive generalization is the following: the vowel of the prefix is [ɛ] if the first vowel of the lexical item is [ɛ, i] and [ɔ] if the vowel is [ɑ, ɔ, u]. For example, a lexical item such as [ˈlinɔ] has a prefixed form [ɛnˈlinɔ] with a front vowel in the prefix [ɛn-] because the first vowel in the lexical item [i] is front. A lexical item such as [ˈluru] has a prefixed form with [ɔn-]: [ɔnˈluru] because the first vowel of the lexical item [u] is not front. The experiment thus features a similar case of vowel harmony as the Turkish example (see Section 1).

The computational experiment presented here tests the learning of non-local vowel harmony. That the process tested here is phonologically non-local is clear from Table 1: the sounds in correspondence (the vowel of the prefix and the first vowel of the lexical item) are always separated by one or two consonants.

2.2.2. Local processes

In addition to non-local vowel harmony, the training data contain evidence for four local processes that are triggered by the prefix. Two processes are triggered by a nasal sound in the prefix VN-. 16 unprefixed-prefixed pairs (32 items total) contain evidence for post-nasal devoicing (D → T / N____), where a voiced stop devoices if a nasal precedes it: [ˈ**b**ɑlu] ∼ [ɔmˈ**pʰ**ɑlu]. In another 16

[1] Radford et al. (2015) uses averaging over z-variables in some cases and performs logistic regression on the second to last convolutional layer.

pairs (32 items total), a voiced fricative gets devoiced and occluded when a nasal precedes it (Z → T / N___): [ˈ**v**iɹə] ~ [ɛmˈ**pʰ**iɹə]. The other two processes are triggered by the V-prefix. The evidence for intervocalic devoicing, where voiced stops devoice intervocalically (D → T / V___V) is present in 16 unprefixed-prefixed pairs (32 items total), e.g. [ˈ**b**ulɔ] ~ [ɔˈ**p**ʰulɔ]. Another 16 pairs (32 items total) contain evidence for intervocalic fricativization and devoicing, where voiced stops fricativize and devoice (D → S / V___V) between vowels (triggered by the prefix), e.g. [**b**ɔɹə] ~ [ɔˈ**f**ɔɹə]. In the 54 remaining pairs (108 total), no consonantal changes are present, e.g. [ˈjɑlu] ~ [ɔˈjɑlu] or [ˈɹinu] ~ [ɛnˈɹinu].

Because the learning of non-local processes is predicted to be more difficult than that of local processes, the training data contain substantially more evidence for the non-local process. All items in which C_1 is constant as well as those in which it changes contain evidence for the non-local vowel harmony process. Of 270 training items, there are 117 unprefixed items with 117 corresponding prefixed forms, all of which contain evidence for vowel harmony (234 total). The remaining items (36) only include unprefixed forms (for testing learning). There is thus a substantial difference in the amount of training data that contain evidence for the non-local process (117 pairs, 234 altogether) and the four local processes (16 pairs each). Even if all four local processes are pooled together, the data still contain only 64 pairs containing evidence for the four local processes (128 altogether). Table 1 illustrates the training data: each slot is filled with a transcribed example from the training data. The entire training in IPA transcription is given in Appendix Tables A.3, A.4, A.5, A.6, A.7, A.8, and A.9.

In addition to the local and non-local processes described above, the data contain evidence for a local assimilation process which is somewhat less relevant to our experiment: if the prefix contains a nasal stop (VN-), the place of articulation of the nasal stop depends on the first consonant of the root (C_1). The nasal surfaces as labial [m] before the labials ([p] and [f]), and as an alveolar [n] elsewhere. Spectral differences are minimal between the two conditions, which is why a detailed analysis of this process is not possible in the computational experiment; the main purpose for including this assimilation in the data is for the behavioral experiment to include an English-like process (to not raise the attention of the subjects) and to facilitate the reading task for the speaker who recorded the stimuli.

The computational experiment tests the learning of the local devoicing processes and non-local vowel harmony that target the PREFIX (VN- or V-). In order to control for the potential effects of other segments on the learning of the targeted processes, we balance the experimental design as much as possible. The number of lexical items with the front vowel in V_2 is, in all but three pairs, equivalent for every C_1 condition. In other words, if there are four [d]-initial items that devoice and have frontness harmony (V_2 is front), there are also four items with backness harmony (V_2 is not front) for this condition.[2] We also aim to balance the identity of C_3 and V_4 as much as possible, but balancing these positions is limited by the requirement that the items not be real words of English or too similar to real words (due to the artificial grammar learning experiment). Only [m, n, l, ɹ, s] can be members of C_3, and these along with V_4 are relatively well balanced across the groups with changing C_1 (e.g. approximately equal number of the same consonants across voiced-initial items that devoice and those that undergo devoicing with fricativization or occlusion), but not across other groups. A fully balanced design is difficult to achieve due to different groups and the nonce-word requirement, but given the relatively well balanced design, we do not expect undesired dependencies to affect the learning distributions of interest.

[2] There are two missing frontness harmony pairs in the non-changing [pʰ]- and [tʰ]-initial condition and one missing backness harmony pair in the non-changing [l]-initial condition for the VN- prefix.

Table 1: Examples of words used in training in the IPA transcription.

Prefix			**Labial**		**Coronal**		**[j]**	**[l]**	**[ɹ]**
VN-	**C$_1$ constant**	**ɛ-harmony**	ˈp^{h}imi	ˈfimə	ˈt^{h}ɛlɔ	ˈsɛnɔ	ˈjim	ˈlɛn	ˈɹinu
			ɛmˈp^{h}imi	ɛmˈfimə	ɛnˈt^{h}ɛlɔ	ɛnˈsɛnɔ	ɛnˈjim	ɛnˈlɛn	ɛnˈɹinu
		ɔ-harmony	ˈp^{h}ɔɹɔ	ˈfuɹə	ˈt^{h}ɑɹu	ˈsɑnu	ˈjɑlu	ˈlɔɹ	ˈɹɔlɔ
			ɔmˈp^{h}ɔɹɔ	ɔmˈfuɹə	ɔnˈt^{h}ɑɹu	ɔnˈsɑnu	ɔnˈjɑlu	ɔnˈlɔɹ	ɔnˈɹɔlɔ
	C$_1$ changes	**ɛ-harmony**	ˈbeɹə	ˈvirə	ˈdɛlɔ	ˈziɹə	—	—	—
			ɛmˈp^{h}eɹə	ɛmˈp^{h}irə	ɛnˈt^{h}ɛlɔ	ɛnˈt^{h}iɹə	—	—	—
		ɔ-harmony	ˈbɑlu	ˈvɔnə	ˈdunə	ˈzɔlɛ	—	—	—
			ɔmˈp^{h}ɑlu	ɔmˈp^{h}ɔnə	ɔnˈt^{h}unə	ɔnˈt^{h}ɔlɛ	—	—	—
V-	**C$_1$ constant**	**ɛ-harmony**	ˈp^{h}inə	ˈfini	ˈt^{h}ɛlɔ	ˈsɛnɔ	ˈjim	ˈlinɔ	ˈɹɛl
			ɛˈp^{h}inə	ɛˈfini	ɛˈt^{h}ɛlɔ	ɛˈsɛnɔ	ɛˈjim	ɛˈlinɔ	ɛˈɹɛl
		ɔ-harmony	ˈp^{h}ɔmɔ	ˈfuɹə	ˈt^{h}ɔmɔ	ˈsɑnu	ˈjɑm	ˈluɹu	ˈɹɑs
			ɔˈp^{h}ɔmɔ	ɔˈfuɹə	ɔˈt^{h}ɔmɔ	ɔˈsɑnu	ɔˈjɑm	ɔˈluɹu	ɔˈɹɑs
	C$_1$ changes	**ɛ-harmony**	ˈbɛlə	ˈbɛmə	ˈdɛni	ˈdɛmɛ	—	—	—
			ɛˈp^{h}ɛlə	ɛˈfɛmə	ɛˈt^{h}ɛni	ɛsɛmɛ	—	—	—
		ɔ-harmony	ˈbulɔ	ˈbɔɹə	ˈdɑɹu	ˈdɑlə	—	—	—
			ɔˈp^{h}ulɔ	ɔˈfɔɹə	ɔˈt^{h}ɑɹu	ɔˈsɑlə	—	—	—

The 270 items described above were presented in a simplified transcription (see Appendix Tables A.3, A.4, A.5, A.6, A.7, A.8) and read by a single female speaker of American English. The words were of the shape C$_1$V$_2$C$_3$, C$_1$V$_2$C$_3$V$_4$, PREFIX-C$_1$V$_2$C$_3$, and PREFIX-C$_1$V$_2$C$_3$V$_4$. The prefixes were of the shape VN- and V-: [ɛn-], [ɔn-], [ɛm-], [ɔm-], [ɛ-], and [ɔ-] (see Appendix A Tables A.3, A.5, and A.7). The speaker was unaware of the exact objectives and details of the study and was compensated for her work. Recordings of training data were made in a sound-attenuated booth using a USBPre 2 (Sound Devices) pre-amp and Shure 53 Beta omnidirectional condenser head-mounted microphone in Audacity (originally sampled at 44.1 kHz and then downsampled to 16 kHz).

The data in the form of sliced audio files for each item (approximately 1 s long padded with silence) is fed to the model randomly in mini-batches of 64. The bare unprefixed and prefixed forms are not paired in any way during training.

3. Results

One advantage of the GAN architecture is that the Generator network outputs innovative data that are linguistically interpretable (Beguš, 2020c). Innovative outputs are often sporadic and do not allow for a full quantitative analysis, which nonetheless does not make them less informative. It is important to describe innovative outputs and how they can inform us about the learning of speech data in deep convolutional networks. In Sections 3.1 and 3.2 we present results from an exploratory study of the network's innovative outputs based on an acoustic analysis of spectra. In Sections 3.3, 3.4, and 3.5 we present a quantitative analysis of the generated outputs.

3.1. Small data sets

The total unique data points (audio recordings of the words with the structure described in Section 2.2) that the network is trained on is 270. Despite the small amount of training data, the model generates outputs that closely resemble human speech, are interpretable, analyzable, and highly informative. This stands in contrast to some recent studies of neural network models on the syntactic level that require very large training datasets and do not improve substantially with more data (van Schijndel et al., 2019). As is argued below, the GANs do not overfit, but produce innovative data that are linguistically interpretable despite the small training data set. This finding should open up numerous possibilities for further exploration of learning representations in deep convolutional networks: it is generally assumed that GANs and deep convolutional networks require large amounts of data, which could be prohibitive for research questions that require smaller training datasets.

We analyze outputs of the Generator network at four training steps: after 7453 (∼ 8833 epochs), 9740 (∼ 11543 epochs), 14900 (∼ 17659 epochs), and 20990 (∼ 24877 epochs) steps. The number of steps chosen is based on maximizing clarity of the acoustic outputs that need to be appropriate for acoustic analysis and minimizing the number of steps used for training (for guidelines, see Beguš 2020c).

Some generated outputs are phonetically very similar to the input equivalents, as illustrated in Appendix A Figure A.11. The network, however, also generates outputs that substantially violate the input data. The Generator network trained after 7453 steps, for example, outputs a sequence that can be transcribed as [ˈdinɔ], yet the training data lacks this sequence altogether. The closest neighbor to the innovative [ˈdinɔ] in the training data is [ˈdɛnɔ] (see Figure A.11). There are numerous other such generated outputs that violate the training data, but are linguistically valid and interpretable. For example, 23.2% of outputs violate the training data with respect to vowel harmony (see Section 3.4). Innovative outputs that violate training data distributions in linguistically interpretable ways constitute strong evidence against overfitting in the GAN architecture: even with very small datasets and a relatively high number of epochs, the Generator does not overfit. This is in line with previous evidence that GANs generally do not overfit (Adlam et al., 2019; Donahue et al., 2019), but here we additionally argue that GANs don't overfit even with small training datasets (N = 270).

3.2. Progression of learning

One advantage of the exploratory study of GANs outputs is that we can follow how dependencies in speech are learned by the network at different training steps. We propose that the progression of learning can be observed by keeping the latent space constant and generating data at different training stages of the Generator network. This provides crucial information on how the number of training steps influences the Generator's outputs and learning representations — an area that is relatively understudied. Testing the effect of training steps on learning representations using speech data should reveal further insights into neural network interpretability, as is argued below.

We propose that by analyzing generated outputs at different training steps with latent space kept constant, we can actively follow how the network corrects the outputs that violate distributions in the data. For example, at 7453 steps, the network generates an innovative output that violates the training data: [ˈbɛnɔ]. At 9740 training steps, the network outputs [ˈbɛmɔ] for the same latent space variables. This output still violates the data: none of the words in the training data was of the exact shape [ˈbɛmɔ]. At 14900 steps, the network outputs [ˈbɛɹɔ] (for the same latent space), which corresponds to [ˈbɛɹɔ] in the training data (Figure 2).

In a related example, the proposed method allows us to follow how the network searches through the space of possible segment combinations using linguistically valid strategies. Figure 2 shows an output [ˈzilɔ] for which there is no direct equivalent in the training data. The spectrogram shows a clear voicing bar and frication noise in the high frequencies, characteristic of a [z]. At 9740 steps, the network devoices the initial consonant C$_1$, but keeps its frication noise (and also changes the high front vowel [i] to a back vowel [u] for an output [ˈsulɔ]. This output is likewise not attested in the training data. Finally, at 14900 steps, the network transforms the frication noise from a higher to lower kurtosis that corresponds to a labial fricative [f] in the training data ([ˈfulɔ]). At 20990 steps, it appears as if the network is introducing a period of aspiration noise and turning the fricative into a stop with the same following sequence [ˈt$^{\mathrm{h}}$ulɔ]. None of these outputs are attested in the training

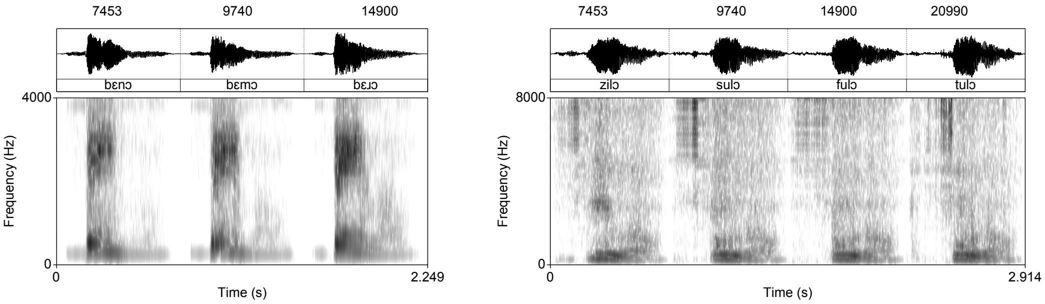


Figure 2: **(left)** Three generated samples with the same values of latent space variables. **(right)** Four generated samples with the same values of latent space variables at four training steps showing devoicing change of place of articulation, and occlusion.

data, but the example illustrates that the Generator searches for segment combinations with valid phonological processes in human language, such as *devoicing* or changing *place of articulation.*

Using this technique, we can not only observe how the network repairs distributional violations, but also how it searches through the space of possible segment combinations to repair violations of phonological rules in the data. Because the error rate of local phonological processes is relatively low in the output data, (1.8% at 20990 steps), the study of how the network repairs outputs that violate phonological processes can only be exploratory at this point. An example that illustrates how learning progress can be directly observed with this method is given in Figure 3. At 7453 training steps, the Generator outputs [ɛˈzɑɹɔ] which violates both the local process of devoicing after a prefix and the non-local vowel harmony process. At 9740 steps, the second formant of the prefix vowel ([ɛ]) substantially weakens and the formant structure of a back [ɔ] emerges, which means the network repairs the harmony violation. At 14900 steps, voicing in the fricative ceases from the output, which means the output now conforms to the devoicing rule in the training data. In other words, [z], which violates the phonological rule of devoicing after a prefix, devoices to [s], which conforms to the training data. At 14900 steps, the output thus fully conforms to the distributions in the training data: harmony and devoicing: [ɔˈsɔlɔ] (Figure 3). The output, while conforming to the rules of training data, is still innovative and none of the training inputs contains exactly this sequence. Spectrograms in Figure 3 illustrate how the network applies learning representations in its continuous outputs at different training steps that correspond to phonological processes in natural language: *devoicing* and *vowel-lowering.*

3.3. Latent space

To test how the network encodes prefixation in its latent space, we used a technique described in Beguš (2020c) and Section 4 to identify dependencies between the latent space and generated data. 500 outputs of the Generator network trained after 20990 steps were transcribed and annotated for presence of the prefix V- and VN-.[3] The number of steps for this analysis was chosen based on the analysis of progression of learning in Section 3.2: it appears that a number of disharmonic outputs is repaired at 20990 steps and further training with more steps ceases to repair disharmonic outputs. That the network is successful in outputting data that approximates human speech in the training data is suggested by the fact that the author was unable to reliably transcribe the output in only 25 out of 500 outputs (5%). The data were fit to a Lasso logistic regression model with the presence of the prefix as the dependent variable and the 100 latent variables of the Generator network as

[3] All acoustic analyses are performed in Praat (Boersma and Weenink, 2015).

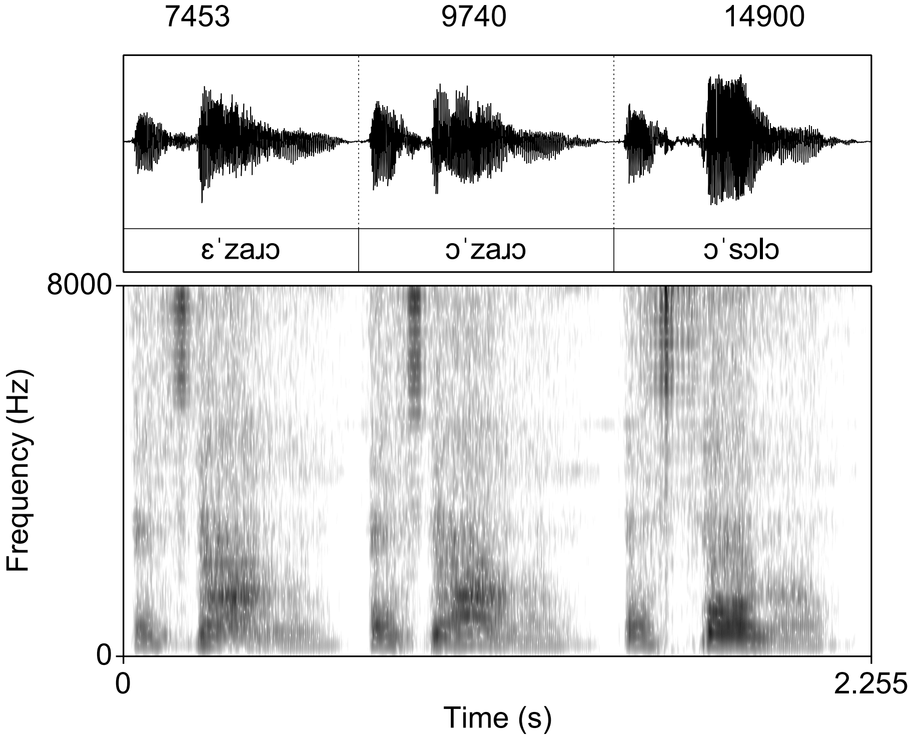


Figure 3: Changes in outputs with the same latent space values across training steps.

predictors (with the *glmnet* package in Simon et al. 2011). Alpha values were estimated with 10-fold cross-validation. Estimates in Figure 4 suggest that the network uses a single latent variable to encode the presence of the prefix in the output: there is a clear and substantial drop in regression estimates between z_{16} and the rest of the latent space (other 99 z-variables). Such a substantial drop in regression estimates suggests that the network discretizes representation of the prefix into a single latent variable.

To test the effect of z_{16} on generated data, we generate 100 outputs with the value of z_{16} set at -4.5 (for the method, see Beguš 2020c and Section 2.1). Out of 100 generated samples, 100 (or 100%) contain a prefix V- or VN-. When z_{16} is set to its opposite value (4.5), only 1 out of 100 generated samples (1%) contains a prefix. This generative test suggests that the network encodes presence of the prefix in the output as a single variable in its latent space. By manipulating this feature, we can actively control the presence of the prefix in the output.[4]

3.4. Local and non-local processes

The training data contains evidence for local and non-local phenomena. Devoicing and occlusion after the prefixes V- and VN- are local; vowel harmony is non-local, as one or two segments intervene between the target and the corresponding vowel.

To test error rates of the output data, 500 outputs from the Generator networks trained after 20990 steps were analyzed. 211 outputs (42.2%) were analyzed as involving a prefix VN- or V-. Of

[4] For a generative test showing that regression estimates indeed identify variables that correspond to a given phonetic/phonological representation, see Beguš (2020c).

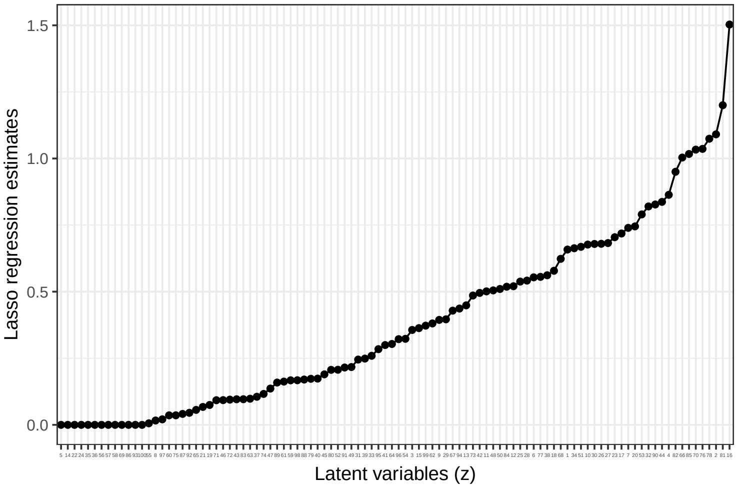


Figure 4: Absolute Lasso logistic regression estimates of a model with presence of the prefix as the dependent variable and values of 100 z-variables as independent predictors). The estimates are sorted in reversed order.

	VN-		**V-**		**Total**
	front	**back**	**front**	**back**	
Harmonious	53	31	47	31	162
Non-harmonious	21	6	15	6	48
% Harmonious	71.6%	83.8%	75.8%	83.8%	77.1%

Table 2: Raw counts of harmonious and disharmonious outputs of the Generator network across the two prefixes and vowel quality levels (front vs. back).

the 211 prefixed outputs, 162 (or 76.8%) were analyzed as harmonious.[5] Harmonious responses are consistently more frequent than non-harmonious both for front and back V_2 as well as across the two prefixes, V- and VN-. The distribution of the harmonious and disharmonious outputs across front and back triggering vowels and across the two prefixes are given in Table 2.

To test whether the Generator's higher rates of harmonious responses are significantly above chance, we fit the data to a linear logistic regression model with harmonious responses as a dependent variable (coded as successes) and vowel FRONTNESSS (with two sum-coded levels, front and back) and PREFIX identity (with two sum-coded levels, V- and VN-) as the independent variables with their interaction. Harmonious responses are significantly more frequent than disharmonious responses at means of all predictors: $\beta = 1.34, z = 7.2, p < 0.0001$. None of the interactions are significant. All estimates are given in Appendix Table A.10. Predicted values of the model are plotted in Figure 10. The results suggest that the network learns the non-local phonological process of vowel harmony, but imperfectly so: it violates the training data in approximately 24% of outputs. The violations are linguistically interpretable: the prefix vowel in the non-harmonious condition is not of random formant structure, but consists of formants characteristic of [ɔ].

Local processes are substantially less frequent and easier to learn than non-local processes in natural languages. To test whether such distribution also emerges in deep convolutional networks, we can compare the error rate in the non-local process and the error rate in the local processes of the generated outputs. Out of 168 prefixed outputs containing a stop or a fricative, only three (1.8%)

[5] In one output excluded from the analysis, the prefix vowel is analyzed as [ɑ].

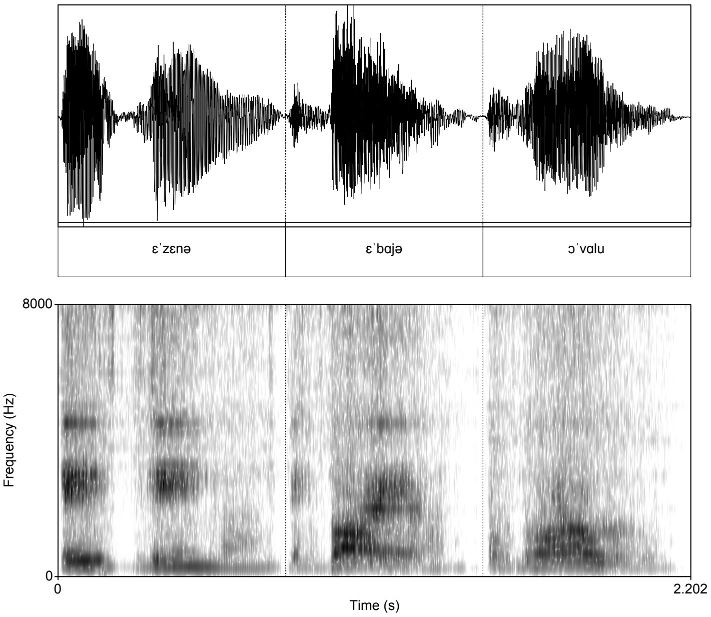


Figure 5: Waveforms and spectrograms (0–8000 Hz) of three outputs of the Generator network trained after 20990 steps, [ɛˈ**z**ɛnə], [ɛˈ**b**ɑjə], and [ɔˈ**v**ɑlu], that violate the training data distributions with respect to local processes of fricative and stop devoicing.

violate the devoicing rule in the training data by which stops and fricatives are always voiceless in prefixed forms, e.g. [ɛˈ**z**ɛnə], [ɛˈ**b**ɑjə], and [ɔˈ**v**ɑlu] (spectrograms in Figure 5). This error rate is significantly lower compared to the error rate of the non-local process (OR =16.5 [5.2, 84.6], $p < 0.0001$, Fisher Test). While the phonetic cues for harmony and devoicing are different and challenging to compare, it would be difficult to argue that the magnitude of phonetic cues for vowel formants (front vs. back) is substantially smaller than the cue for voicing. The distribution aligns well with behavioral data in human subjects, where local processes have been shown to be easier to learn than non-local processes in many studies (Finley, 2011, 2012; McMullin and Hansson, 2019; White et al., 2018).

3.5. Emergence of rule-like behavior

In the framework of symbolic representations, vowel harmony can be derived with an algebraic rule (as in 1). The harmony of the prefix vowel ([ɛ]/[ɔ]) is triggered by the following vowel V_2: a rule that sets the feature [±front] in the vowel of the prefix according to the value of the same feature in the following vowel (see formalism in 1). Alternatively, the grammar can also operate on a morphophonological level: a prefix as a morphological unit can be chosen based on the value of the following vowel.

We propose here that using the technique in Beguš (2020c), we can elicit such rule-like behavior in deep convolutional neural networks. The analysis in Section 3.3 suggests that the Generator learns to associate z_{16} with presence of a prefix. There is a substantial drop in regression estimates after the estimates for z_{16}, which suggests that the network discretizes the continuous phonetic input and uses a single variable to encode presence of some phonetic/phonological material which corresponds to a morphological unit: a prefix. To elicit rule-like behavior, we can identify another variable in the latent space — the variable that corresponds to the frontness/backness of vowel V_2. To identify such a variable, we generated 500 outputs are annotated the outputs for vowel (V_2)

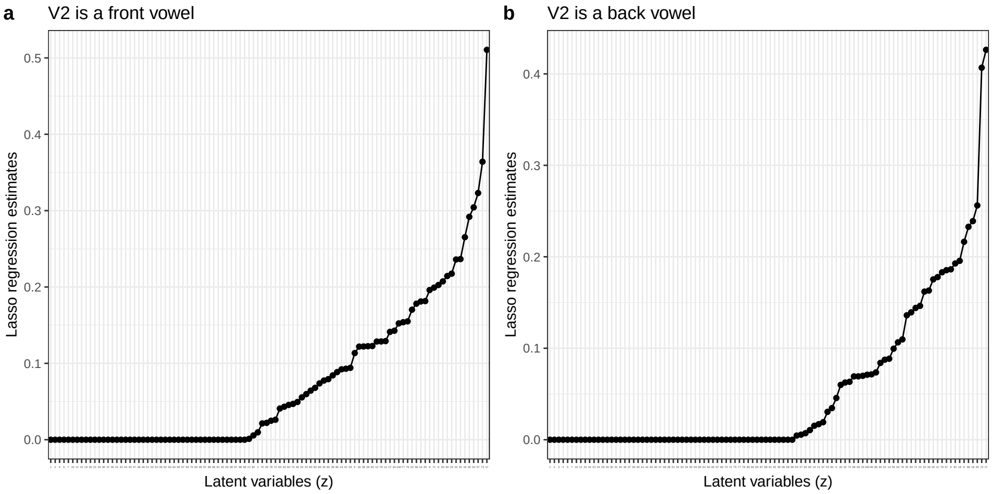


Figure 6: **(a)** Absolute Lasso logistic regression estimates of a model with presence of front triggering vowels V_2 as the dependent variable and values of 100 z-variables as independent predictors). The estimates are sorted in reversed order. **(b)** Absolute Lasso logistic regression estimates of a model with presence of back triggering vowels V_2 as the dependent variable and values of 100 z-variables as independent predictors). The estimates are sorted in reversed order.

frontness. We fit the data to two linear logistic regression models: one in which outputs with the front vowel (V_2) [ɛ, i] are coded as success and another in which [ɑ, ɔ, u] are coded as success. The independent variables are values of the 100 latent variables z randomly sampled for each of the 500 annotated generated outputs. The model is fit using the *glmnet* package (Simon et al., 2011) in R (R Core Team, 2018). Lambda values are estimated with 10-fold cross-validation. Estimates of the two models are given in Figure 6.

Both models uniformly suggest that z_{17} is the latent variable most strongly associated with determining vowel frontness of the triggering vowel V_2. Regression estimates again suggest that the Generator network learns to encode vowel frontness with a single latent variable: there is a substantial drop of estimates after the single latent variable z_{17}. Negative values of z_{17} correspond to presence of front [ɛ, i] in V_2, while positive values correspond to presence of back [ɑ, ɔ, u] (estimates in Figure 6 are in absolute values).

To elicit rule-like behavior, we force the prefix in the input and simultaneously force vowel V_2 to turn from a front vowel [ɛ, i] into a back vowel [ɑ, ɔ, u]. To achieve this affect, we simultaneously manipulate z_{16} (presence of prefix) and z_{17} (frontness of vowel).[6] If the Generator network learned vowel harmony, then the vowel of the prefix should change together with the forced change of vowel quality. Such a behavior would parallel rule-based computation: setting a single variable to a value that forces prefixation in the output and manipulating the variable that changes the conditioning environment(V_2) results in a process that changes the target vowel according to the condition — vowel harmony.

To test this hypothesis, we set the value of z_{16} to -2.5 which forces the prefix in the output. Additionally, we generate outputs with z_{17} interpolated from values -6 to 6 in increments of 1. 60 such sets of 13 generated samples (with z_{17} from -6 to 6) are generated and acoustically analyzed (780 outputs total). That z_{16} indeed causes the prefix in the output is suggested by the count of prefixed forms in the output: 634 out of 780 generated samples (or 81.3%) were analyzed as

[6] That the two variables are consecutive is coincidence.

featuring a prefix (for an independent test of the effect of z_{16} on presence of prefix, see Section 3.3).

That z_{17} indeed changes the triggering vowel V$_2$ from a front [ɛ, i] to a back [ɑ, ɔ, u] is strongly suggested by the generated outputs. We annotate the 634 prefixed forms from the 60 sets of generated interpolated outputs for frontness and backness of the triggering vowel V$_2$. We fit the annotated data to a generalized additive mixed logistic regression model (GAMMs; Wood 2011) with an intercept and thin-plate smooths that estimate how the presence of a front or back vowel in the output changes with interpolated values. A random smooth for each trajectory (each of the 60 generated sets) is added to the model. Figure 7 suggests that the presence of z_{17} causes the triggering vowel from a front one at values in the negative range to a back one at positive values. The relationship appears to be linear even when the model does not have an assumption of linearity (GAMM). If we refit the data to a linear logistic mixed effect regression (with a random intercept for trajectory and by-trajectory random slopes), we get a significant negative correlation between values of z_{17} (from -6 to 6) and percent of front vs. back response ($\beta = -1.01, z = -5.50, p < 0.0001$). Figure 7 illustrates how rates of front vowel V$_2$ in the output change from almost 100% at one end of spectrum to 0% (or 100% of back vowel) in the other end of spectrum.

To test whether the prefix vowel is harmonious even when the variable changing the triggering vowel is interpolated, we annotate the 634 prefixed forms from the 60 sets for frontness of the triggering vowel V$_2$ and for vowel harmony. Data is annotated for harmony (successes vs. failures) and fit to a generalized additive mixed effects logistic regression model. The independent variables are FRONTNESS of the vowel (treatment-coded with back as reference) and a thin plate smooths for values of z_{16} as well as by-trajectory random smooths. The estimates of the parametric term suggest that the prefix vowel is harmonious both for front and back triggering vowels V$_2$. Harmonious outputs with a back triggering vowel V$_2$ ([ɑ, ɔ, u]) are significantly more frequent that non-harmonious outputs: $\beta = 1.35, z = 4.12, p < 0.0001$. That the same is true for the front vowel is clear from estimates in Figure 7 (confidence intervals do not cross zero) and from the fact that estimates for the front triggering vowel V_2 are not different from estimates for back vowel. This is confirmed if we refit the model with sum-coded FRONTNESS factor ($\beta = 1.35, z = 6.23, p < 0.0001$). We also observe a slight negative trend in harmonious responses as we increase z_{17} and a slight positive trend for harmony in the back vowel conditions, although estimates for smooths are not significant. This likely results from the trend that we observe in the data: as we force the triggering vowel away from front (by setting z_{17} to -6), the prefix is harmonious. When the vowel changes as we interpolate the value of z_{17}, we have a higher proportion of disharmonious outputs, because apparently the underlying value of the triggering vowel is not "strongly" front or back. As the value of z_{17} increases towards 6 and the back vowel is forced more strongly in the output, we get a higher proportion of harmonious outputs again (of course with a back vowel harmony).[7] Figure 8 illustrates the gradual change of the forced prefix from a front (containing an [ɛ]) to back (containing an [ɔ]) when z_{17} changes the vowel V$_2$ from a front to a back vowel. In other words, as we force a change of the triggering vowel quality from front to back with a single latent variable, the prefix (also forced with a single variable) automatically changes in order to remain harmonious.

The deep convolutional network thus appears to represent what would approximate a rule-like computation in phonology: as we force the prefix in the output and change the quality of the triggering vowel from front to back by manipulating only two latent variables, vowel harmony emerges automatically. While the appearance of rule-based computation is not categorical — but as is always the case in connectionism, probabilistic — and other features can change along the observed

[7] While the estimates of the effects are significant, the trends are not categorical. Occasionally, the vowel does not change from front to back and more rarely, trends are reversed.

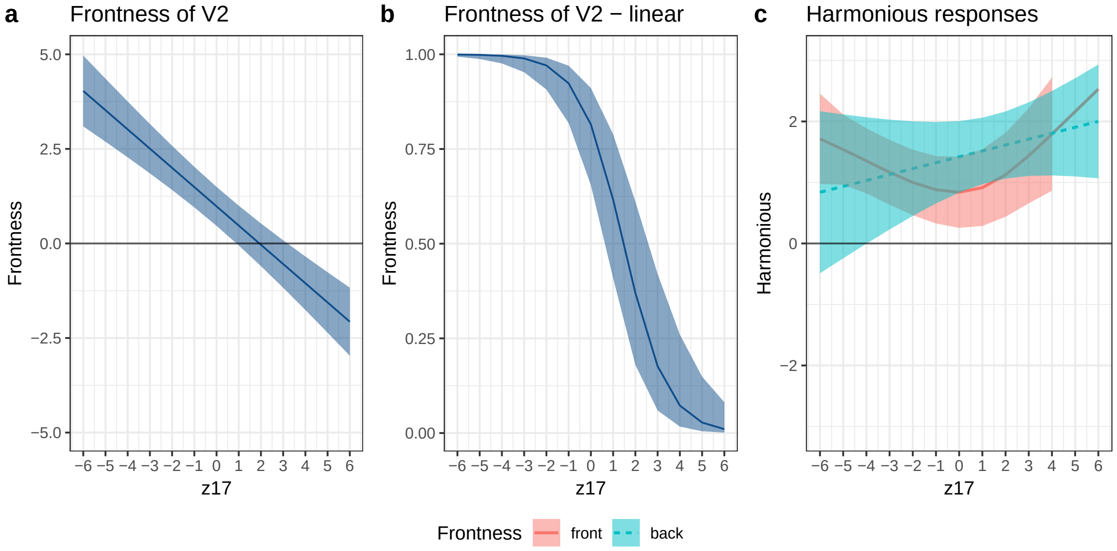


Figure 7: **(a)** Fitted values and 95% CIs of a generalized additive mixed effects logistic regression model with the front vs. back triggering vowel (V_2) value as the dependent variable and thin-plate smooths for values of z_{17} as the independent variable (with random smooths for each of the 60 generated sets). The estimates show that z_{17} causes a change from a front to a back vowel as its values are interpolated from -6 to 6 and that the relationship between values of z_{17} and frontness/backness of the vowel are linear. The regression estimates are in Appendix Table A.12. **(b)** Fitted values and 95% CIs of a linear mixed effects logistic regression model with the front vs. back triggering vowel (V_2) value and random intercept for each of the 60 trajectories. The plot illustrates how the percent of front vowels decreases as the value of z_{17} increases (and vice-versa for back vowels). **(c)** Fitted values and 95% CIs of a generalized additive mixed effects logistic regression model with harmonious (success) and disharmonious (failure) outcome as the dependent variable, vowel FRONTNESS as a parametric predictor, and thin-plate smooths for the two levels of frontness (front vs. back) across the values of z_{17} and random smooths for each of the 60 set of generated outputs. Estimates of the model are given in Appendix Table A.13.

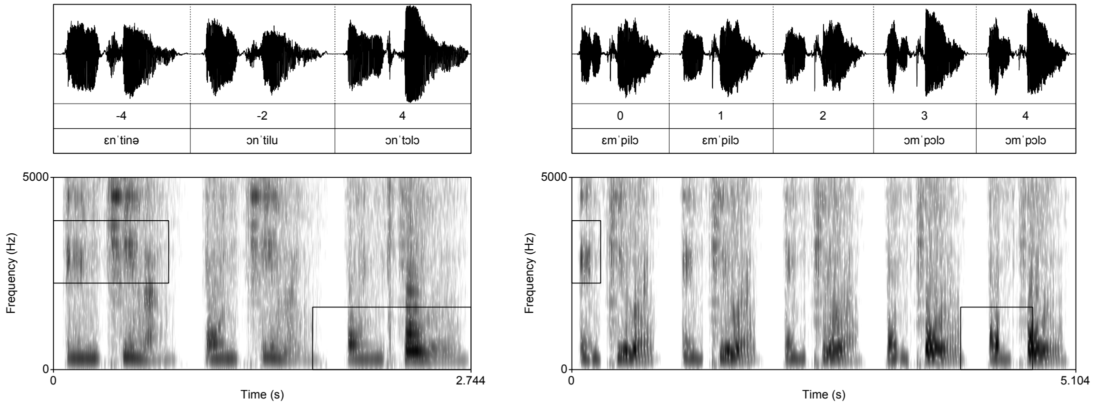
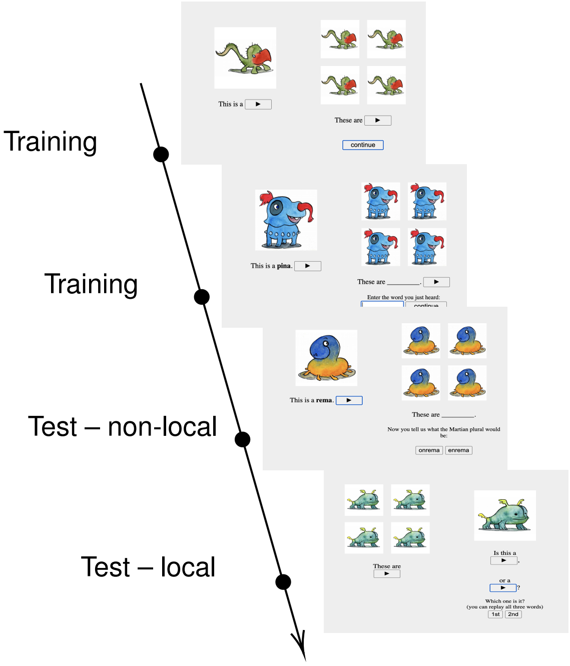


Figure 8: Outputs with interpolated values of z_{17} that change the triggering vowel V_2 from a front [ɛ, i] to a back [ɑ, ɔ, u] and z_{16} set at -2.5, which forces the prefix in the output. **(left)** Three outputs with z_{17} set at $-4, -2$, and 2. The spectrogram shows how the formant structure of a front [ɛ] in the prefix changes to the formant structure of a back [ɔ] as the triggering vowel changes from a front [i] to a back ɔ. **(right)** Five outputs with z_{17} set at $0, 1, 2, 3$, and 4. The spectrogram again shows an automatic change of the prefix vowel consistent with the vowel harmony in the training data. Areas in squares indicate formant structures of interest.



This is a ▶

These are ▶

This is a **pina**. ▶

These are ________. ▶

Enter the word you just heard:

This is a **rema**. ▶

These are ________.

Now you tell us what the Martian plural would be:

These are ▶

Is this a ▶,

or a ▶?

Which one is it?
(you can replay all three words)

Figure 9: Experimenal design (from Beguš 2020b) in the Experigen interface (Becker and Levine, 2013). The order of the training and the test phases are randomized, but the training precedes the test block . For the exact procedure of the experiment, see Beguš 2020b

changes. This is to the author's knowledge the closest approximation of rule-based phenomena, especially considering that the models contain no language-specific mechanism and are trained in an unsupervised manner from raw acoustic data.

4. Paralleling neural networks and artificial grammar learning experiments

To parallel the performance of the computational experiment with results from a behavioral experiment, we combine novel data presented here for the first time with results of an experiment in Beguš (2020b). The subjects were trained on the same data as used in the computational experiment, but divided into two separate experiments: one in which subjects were trained on data with the VN- prefix and another one on data with the V- prefix. Subjects were recruited via Amazon MTurk[8], completed informed consent before participating, and were presented with experimental stimuli in Experigen (Becker and Levine, 2013). In the behavioral experiment, the unprefixed-prefixed forms are presented to subjects in pairs, where the prefixed form carries the function of plural. Subjects were presented with a picture of a Martian creature. A single creature is associated with the unprefixed form; four creatures are associated with the prefixed form. The experimental interface is illustrated in Figure 9.

Subjects whose first language was not English or who had self-reported linguistic education

[8] That the results of the experiment are not heavily influenced by the fact that the participants in the behavioral experiments are recruited via Amazon MTurk is suggested by the fact that vowel harmony outcomes are very similar to a related experiment with very similar training data that wasperformed in-person with the supervision of a research assistant in which subjects were recruited from the general public Beguš (2020b).

were removed from the analysis. Altogether 333 subjects that provided 1987 responses on the vowel harmony test are analyzed [9]

The training phase in the VN- experiment consisted of 58 pairs of bare and prefixed forms. All examples were harmonious and some included evidence for the local processes of post-nasal devoicing and post-nasal devoicing and occlusion (as described in detail in the Section 2.2 on data used in the computational experiment). In the V- experiment, the training phase consisted of 60 pairs of bare and prefixed forms, all of which contained evidence for harmony and some of which contain evidence for local processes of devoicing and devoicing and fricativization (see Section 2.2). All items used in the behavioral experiment are listed in Appendix Tables A.3, A.4, A.5, A.6, A.7, A.8, and A.9.

After the training phase, the subjects were tested on six bare forms with C_1 either a [r] or [l] (three with a front V_2 and three with back) and had to choose between harmonious and non-harmonious responses in a forced choice task (see Test – Local in Figure 9), as well as between various local processes. For example, subjects were presented with a stimulus [ˈlirɔ], presented auditorily and orthographically, and had to choose between the plural form *eliro* (harmonious) and *oliro*, presented only orthographically.[10]

While the behavioral experiments do not directly test whether non-local processes are more difficult to learn than local processes (this has already been confirmed experimentally in several studies; see Finley 2011, 2012; McMullin and Hansson 2019; White et al. 2018), the local process is made more difficult to learn in the experiment: subjects were explicitly instructed to learn the (non-local) distribution of prefixes (vowel harmony), but never about learning the local processes. Moreover, the learning of local processes is tested exclusively with auditory stimuli.

To test the learning of the non-local process in the behavioral experiment, the responses were fit to a linear mixed effects logistic regression model (*lme4* package by Bates et al. 2015). First, we fit the full model with harmonic vs. non-harmonic responses (successes vs. failures) as the dependent variable and FRONTNESS (front vs. back, sum-coded) of the vowel and the shape of the PREFIX (VN- vs. V-, sum-coded) as the independent variable (with interaction) and random intercepts for SUBJECT and ITEM with by-subject and by-item random slope for HARMONY. The final model was chosen based on Akaike Information Criterion (AIC) by removing random slopes first and then interactions. The final model includes the FRONTNESS × PREFIX interaction and random intercepts for SUBJECT and ITEM.

The results show that subject learn the vowel harmony pattern from the training data ($\beta = 0.56, z = 5.0, p < 0.0001$). In other words, harmonious responses are significantly above the chance level, which suggests subjects do learn the harmonious pattern. However, the error rate is quite high. The 95% profile CIs for the preference for harmonious response are quite low: [57.6%, 69.2%], especially given that 234/270 items are bare-prefixed pairs each of which contains evidence for vowel harmony. All regression estimates are in Table A.11.

We can directly compare subject's responses in the behavioral experiments with outputs of the computational experiment. The Generator network violates local distributions in the data in only three out of 168 generated outputs with a prefix and a stop or a fricative (1.8%). On the non-local task, however, the Generator's error rate is substantially higher and similar to the error rate in

[9] For detailed exclusion criteria in the VN-condition, see Beguš (2020b). In the V- condition, we excluded participants with non-unique Amazon MTurk IDs as well as with those IDs who had already taken the VN-experiment.

[10] In the test phase on local processes involving the prefix VN-, the subjects were presented with a plural form exclusively auditorily and had to choose between two possible singular forms: one consistent with devoicing and another consistent with devoicing and occlusion. In the V- condition, the subjects similarly chose between singular forms consistent with intervocalic devoicing or intervocalic devoicing with fricativization.

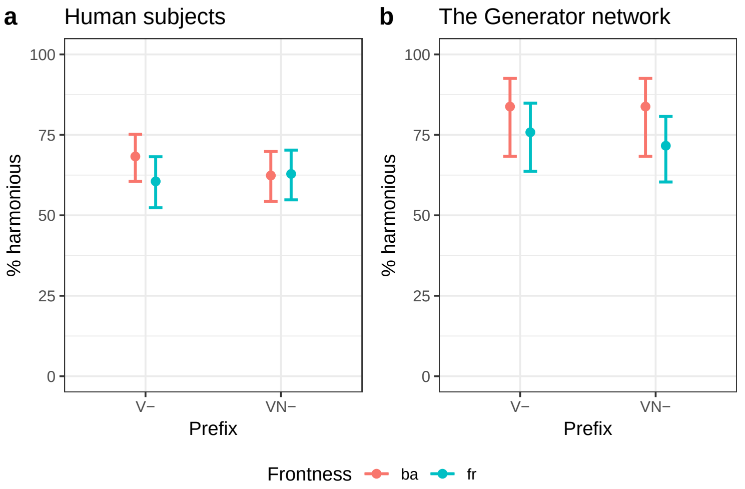


Figure 10: **(a)** Estimates and 95% CIs of the linear logistic regression model with harmonious responses of the Generator network as successes and vowel frontness and prefix identity as the independent variables with their interaction. **(b)** Linear mixed effects logistic regression estimates with harmonious responses of human subjects in the behavioral experiment as successes and vowel frontnesss and prefix identity as the independent variables with their interaction.

the artificial grammar learning experiment conducted on human subjects. Figure 10 illustrates the similarity.

To be sure, there are substantial differences between the computational and behavioral experiment. First, the comparison is necessarily superficial, because this paper does not claim that humans learn phonological patterns in the same way as deep convolutional networks; however, this does not preclude us from comparing their performance. The number of epochs in the computational experiment is $\sim$ 24877, while subjects were only exposed to training data once. On the other hand, human subjects were adults with full language capacity and already established phonological inventories, phonological grammar, and articulatory and perceptual mechanisms. The Generator network has to learn to produce speech-like outputs from random noise and does not contain any language-specific learning mechanisms.

This comparison in performance between human subjects and the computational model suggests that non-local processes are computationally similarly costly both for humans and for computational models of language acquisition to the degree that the error rates across the two conditions are similar. That non-local processes are computationally costly has of course been shown before, but to our knowledge, this is the first such confirmation on a deep convolutional neural network model that is trained on the same data as human subjects and that learns speech representations from raw acoustic data.

5. Discussion

This paper tests learning of local and non-local processes in human speech with deep convolutional networks in the GAN architecture. More specifically, we test the learning of non-local vowel harmony and local devoicing processes in a setting that approximates morphological and phonological processes in language: the model is trained on data with bare and prefixed forms in random order.

First, we argue that deep convolutional GANs output highly informative data despite being trained on extremely small datasets (N = 270) with a high number of epochs. The outputs are acoustically analyzable and linguistically interpretable. As has been shown before (Beguš, 2020c,a), the Generator outputs innovative data that violate training data. These violations, however, are not random, but are linguistically interpretable. 23.2% of outputs are disharmonious, which means they violate training data distributions. But these violations are not random; the network outputs either a front [ɛ] or a back vowel [ɔ] in the prefix, consistent with the shape of the prefix in the training data. In only 5% of annotated outputs is the data not linguistically interpretable. Innovative outputs also suggest that the Generator does not overfit despite the high number of epochs, in line with previous work on overfitting in GANs. The finding that GANs can be trained on very small data sets should open up several new possibilities for research on deep convolutional networks, speech, and internal representations in deep convolutional networks.

An exploratory study of innovative outputs suggests that, in order to repair its data violations, the network uses strategies that approximate processes in human phonology: devoicing, occlusion, and distribution of frication noise. We propose that these repairs can be directly followed with progression of learning by keeping the random variables constant while generating data from the network trained at different training steps. Acoustic analysis of outputs at different training steps in Sections 3.1 and 3.3 identifies strategies that the network uses to repair violations in data distributions.

One of the objectives of this paper is to explore how deep convolutional networks trained in the GAN framework on raw speech discretize linguistically meaningful representations in the latent space, especially with respect to non-local morphophonological processes. The raw acoustic data hearing human infants are faced with is continuous. Phonological computation discretizes the continuous space into discrete representations and manipulates these representations, which results in phonological processes such as vowel harmony. Using the technique in Beguš (2020c), we identify variables in the Generator's latent space that correspond to linguistically meaningful units, such as presence of a prefix or frontness of a vowel. Lasso regression estimates suggest that the network uses a minimal number of variables to represent presence of a prefix in the output. In other words, the steep drop in the regression estimates after the variable with highest estimates suggest that the network discretizes some continuous phonetic content in its internal space. The same is true for a phonetic feature such as frontness of the first vowel in bare forms V_2. The network appears to primarily use a single variable to encode this phonetic property of outputs. An independent generative test suggest that manipulating this one variable on a linear scale well outside the training range (from -6 to 6) results in a gradual and linear transition from a front to a back first vowel in the bare forms (Figure 7).

This paper argues that an approximation of a symbolic rule emerges automatically in deep convolutional networks. To test learning of the non-local vowel harmony, we force a prefix in the output with a single variable (z_{16} at -2.5) and force the change of the triggering vowel from front to back with a linear interpolation of a single variable (z_{17}). The statistical tests in Section 3.5 suggest that the generated output remain harmonious in the majority of cases despite the change of the triggering vowel. In other words, the rule-like vowel harmony emerges automatically in a deep convolutional network from an interaction of the variable that forces some morphophonological entity in the output (the prefix) and the variable that changes the triggering vowel. While harmonic outputs are significantly more frequent than non-harmonic outputs, the distribution is probabilistic rather than categorical. Another trend emerges from the statistical tests: the outputs are more likely to be non-harmonic in the transition period when the triggering vowel changes from front to back. It is likely the case that the relative strength of frontness and backness affects the rates of harmonic vs. non-harmonic outcomes. In other words, it appears that the prefix harmony is not triggered until the frontness/backness feature of the triggering vowel is strong enough, i.e. has a high enough latent variable value. That phonological features bear inherit weights (that can be conceptualized as strength or latent feature values in our model) has been argued before in the

Optimality Theoretic framework (Smolensky and Goldrick, 2016; Smolensky et al., 2019).

Phonological computation has been shown to favor local processes over non-local processes. Many studies show experimentally that the learning of non-local processes is more difficult (Finley, 2011, 2012; McMullin and Hansson, 2019; White et al., 2018). This learning bias is also reflected in typology: the majority of phonological processes are local in the world's languages (Finley, 2011). A clear preference for locality emerges in our computational experiment as well: despite substantially more evidence for the non-local process in the training data, the error rate is significantly higher in the non-local condition in the Generator's network. Whether the prevalence of some patterns in human speech results from articulatory factors (e.g. the articulation of sounds is most strongly affected by the immediately preceding or following sounds) or from learnability (e.g. the learning of non-local processes is more difficult) has been a focal topic of discussion in phonology, linguistics, and cognitive science in general. While this result does not offer an answer as to whether the preference for non-locality in typology results from learning or a language's cultural transmission Beguš (2020b), it does provide evidence that non-locality preferences can be explained with domain-general cognitive mechanisms.

Because GANs trained on small datasets produce informative results, we can use the same stimuli for training deep convolutional networks and artificial grammar learning experiments on human subjects. We compare data from a behavioral experiment that tested the learning of vowel harmony. Results show a similar degree or error rate across the computational and artificial grammar learning experiments. It is true that the Generator network does not output vowel harmony categorically (as opposed to local processes, which are near categorical), but neither do the human subjects tested in a behavioral experiment perform at the categorical level. The training data in the two behavioral experiments contain evidence for vowel harmony almost in every training data point. Harmony is categorical: no training data points violate training data, yet the error rate in human subjects is even higher than in the computational experiments. This suggests that non-local processes are, from a learnability viewpoint, similarly costly both for the deep convolutional network and for human subjects.

Training deep convolutional networks on well-understood dependencies in speech also offers insights into how the network searches through the space of possible segment combinations as the training progresses. We propose that by keeping the latent space constant while generating data from the Generator trained after different number of steps reveals that the networks use linguistically interpretable strategies for repairing the outputs that violate training data. Phonological processes such as devoicing, occlusion, and distribution of frication noise are observed, among others.

The results of the present experiment provide new information on internal representations in deep convolutional networks trained on raw speech, and bear evidence for the long-standing discussion on symbolism vs. connectionism in cognitive science. The networks not only represent morphophonological units with discretized representations (resembling the morphological level), but also learn to encode morphophonological processes (resembling rule-like computation). An approximation of rule-like non-local generalizations in the data emerges from training a deep convolutional GAN. We provide evidence arguing that human behavioral data superficially matches the outcomes of the computational model. Applying such an experiment to further data should yield a clearer picture on how rule-like generalizations emerge as interactions between variables in deep convolutional neural networks trained on raw speech data, and how performance and biases of deep neural networks corresponds to human performance in behavioral experiments.

Acknowledgements

This research was funded by a grant to new faculty at the University of Washington and UC Berkeley as well as by Harvard Mind Brain Behavior and Department of Linguistics.

Adlam, B., Weill, C., Kapoor, A., 2019. Investigating under and overfitting in wasserstein generative adversarial networks.

Alishahi, A., Barking, M., Chrupała, G., Aug. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Association for Computational Linguistics, Vancouver, Canada, pp. 368–378.
URL `https://www.aclweb.org/anthology/K17-1037`

Arjovsky, M., Chintala, S., Bottou, L., 06–11 Aug 2017. Wasserstein generative adversarial networks. In: Precup, D., Teh, Y. W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. Vol. 70 of Proceedings of Machine Learning Research. PMLR, International Convention Centre, Sydney, Australia, pp. 214–223.
URL `http://proceedings.mlr.press/v70/arjovsky17a.html`

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67 (1), 1–48.

Becker, M., Levine, J., 2013. Experigen – an online experiment platform.
URL `http://becker.phonologist.org/experigen`

Beguš, G., 2020a. Ciwgan and fiwgan: Encoding information in acoustic data to model lexical learning with generative adversarial networks, arXiv 2006.02951.
URL `https://arxiv.org/abs/2006.02951`

Beguš, G., 2020b. Distinguishing cognitive from historical influences in phonology, submitted ms., UC Berkeley.

Beguš, G., 2020c. Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. Frontiers in Artificial Intelligence 3, 44.
URL `https://www.frontiersin.org/article/10.3389/frai.2020.00044`

Berent, I., 2013. The phonological mind. Trends in Cognitive Sciences 17 (7), 319 – 327.
URL `http://www.sciencedirect.com/science/article/pii/S1364661313001034`

Boersma, P., Weenink, D., 2015. Praat: doing phonetics by computer [computer program]. version 5.4.06. Retrieved 21 February 2015 from http://www.praat.org/.

Bond, Z. S., Wilson, H. F., 1980. /s/ plus stop clusters in children's speech. Phonetica 37 (3), 149–158.
URL `https://www.karger.com/DOI/10.1159/000259988`

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York.

Donahue, C., McAuley, J. J., Puckette, M. S., 2019. Adversarial audio synthesis. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.
URL `https://openreview.net/forum?id=ByMVTsR5KQ`

Eloff, R., Nortje, A., van Niekerk, B., Govender, A., Nortje, L., Pretorius, A., Biljon, E., van der Westhuizen, E., Staden, L., Kamper, H., 09 2019. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. In: Proc. Interspeech 2019. pp. 1103–1107.

Finley, S., 2011. The privileged status of locality in consonant harmony. Journal of Memory and Language 65 (1), 74 – 83.
URL http://www.sciencedirect.com/science/article/pii/S0749596X11000192

Finley, S., 2012. Testing the limits of long-distance learning: Learning beyond a three-segment window. Cognitive Science 36 (4), 740–756.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2011.01227.x

Garofolo, J. S., Lamel, L., M Fisher, W., Fiscus, J., S. Pallett, D., L. Dahlgren, N., Zue, V., 11 1993. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium.

Gaskell, M., Hare, M., Marslen-Wilson, W. D., 1995. A connectionist model of phonological representation in speech perception. Cognitive Science 19 (4), 407 – 439.
URL http://www.sciencedirect.com/science/article/pii/0364021395900071

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.
URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C., 2017. Improved training of wasserstein gans. In: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5767–5777.
URL http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf

Hansson, G. Ó., 2010. Consonant harmony: Long-distance interactions in phonology. University of California Press.

Heinz, J., 2010. Learning long-distance phonotactics. Linguistic Inquiry 41 (4), 623–661.
URL https://doi.org/10.1162/LING_a_00015

Kabak, B., 2011. Turkish vowel harmony. In: van Oostendorp, M., Ewen, C. J., Hume, E., Rice, K. (Eds.), The Blackwell Companion to Phonology. Wiley Blackwell, Ch. 118, pp. 1–24.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444335262.wbctp0118

Legendre, G., Miyata, Y., Smolensky, P., 1990. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. University of Colorado, Boulder. ICS Technical Report #90-5.

MacMahon, M. K. C., 07 2013. Orthography and the early history of phonetics. In: Allan, K. (Ed.), The Oxford Handbook of the History of Linguistics. Oxford University Press, pp. 105–122.
URL https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199585847.001.0001/oxfordhb-9780199585847-e-6

Marcus, G. F., 2001. The algebraic mind: Integrating connectionism and cognitive science. MIT press.

Marcus, G. F., Vijayan, S., Bandi Rao, S., Vishton, P. M., 1999. Rule learning by seven-month-old infants. Science 283 (5398), 77–80.
URL https://science.sciencemag.org/content/283/5398/77

McClelland, J. L., Elman, J. L., 1986. The trace model of speech perception. Cognitive Psychology 18 (1), 1 – 86.
URL http://www.sciencedirect.com/science/article/pii/0010028586900150

McClelland, J. L., Rumelhart, D. E., , Group, P. R., 1986. Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2. MIT Press, Cambridge, MA.

McMullin, K., Hansson, G., 2019. Inductive learning of locality relations in segmental phonology. Laboratory Phonology: Journal of the Association for Laboratory Phonology 10 (1), 14.

Plaut, D. C., Kello, C. T., 1999. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 381–415.

Prince, A., Smolensky, P., 1993/2004. Optimality Theory: Constraint Interaction in Generative Grammar. Blackwell, Malden, MA, first published in 1993, Tech. Rep. 2, Rutgers University Center for Cognitive Science.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Räsänen, O., Nagamine, T., Mesgarani, N., 08 2016. Analyzing distributional learning of phonemic categories in unsupervised deep neural networks. CogSci ... Annual Conference of the Cognitive Science Society. Cognitive Science Society (U.S.). Conference 2016, 1757–1762.
URL https://pubmed.ncbi.nlm.nih.gov/29359204

Rose, S., Walker, R., 2004. A typology of consonant agreement as correspondence. Language 80 (3), 475–531.
URL http://www.jstor.org/stable/4489721

Rumelhart, D. E., McClelland, J. L., Group, P. R., 1986. Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1. MIT Press, Cambridge, MA.

Shain, C., Elsner, M., Jun. 2019. Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 69–85.
URL https://www.aclweb.org/anthology/N19-1007

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2011. Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software 39 (5), 1–13.
URL http://www.jstatsoft.org/v39/i05/

Smolensky, P., Goldrick, M., 2016. Gradient symbolic representations in grammar: The case of french liaison. In: Rutgers Optimality Archive 1552, Rutgers University.

Smolensky, P., Rosen, E., Goldrick, M., 2019. Learning a gradient grammar of French liaison. In: Proceedings of the 2019 Annual Meeting on Phonology.

van der Hulst, H., 07 2013. Discoverers of the phoneme. In: Allan, K. (Ed.), The Oxford Handbook of the History of Linguistics. Oxford University Press, pp. 167–191.

van Schijndel, M., Mueller, A., Linzen, T., Nov. 2019. Quantity doesn't buy quality syntax with neural language models. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 5831–5837. URL https://www.aclweb.org/anthology/D19-1592

White, J., Nevins, A., Polgárdi, K., Martin, A., Kager, R., Linzen, T., Peperkamp, S., Topintzi, I., Markopoulos, G., van de Vijver, R., 2018. Preference for locality is affected by the prefix/suffix asymmetry. In: Hucklebridge, S., Nelson, M. (Eds.), NELS 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society. GLSA, pp. 207–220.

Wood, S. N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73 (1), 3–36.

Appendix A. Appendix

Appendix A.1. Training data

The recordings or training data were made in a sound-attenuated booth at the Department of Linguistics at Harvard University using a USBPre 2 (Sound Devices) pre-amp and Shure 53 Beta omnidirectional condenser head-mounted microphone in Audacity (originally sampled at 44.1 kHz and then downsampled to 16 kHz).

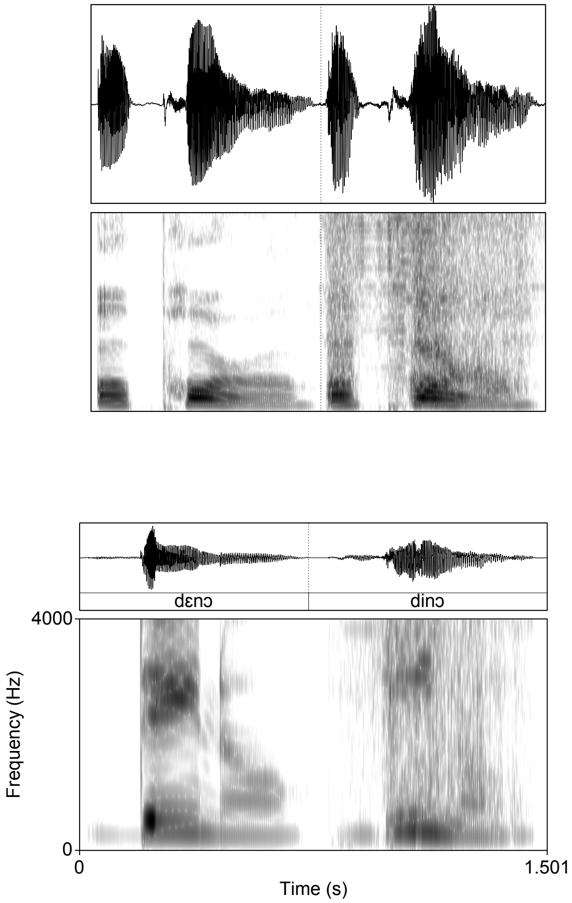


Figure A.11: Waveforms and spectrograms (0–4,000 Hz) of **(top)** input sample (left) and generated sample (right) of [ɔˈp$^{\mathrm{h}}$ɔɾɔ]; and **(bottom)** input sample [ˈdɛnɔ] (left) and generated sample (right) [ˈdinɔ].

Table A.3: IPA transcriptions of training data without consonantal changes; C_1 is a sonorant. [ˈluɹu] and [ɔnˈluɹu] are missing from the computational experiment.

		Fillers			
#__	**Harm.**	**Sg.**	**Pl.**	**Orthography**	
[l]	[+fr]	ˈlɛn	ɛnˈlɛn	len	enlen
		ˈlinɔ	ɛnˈlinɔ	lino	enlino
	[−fr]	ˈlɔɹ	ɔnˈlɔɹ	lor	onlor
		ˈluɹu	ɔnˈluɹu	luru	onluru
[r]	[+fr]	ˈɹɛl	ɛnˈɹɛl	rel	enrel
		ˈɹinu	ɛnˈɹinu	rinu	enrinu
	[−fr]	ˈɹɑs	ɔnˈɹɑs	ras	onras
		ˈɹɔlɔ	ɔnˈɹɔlɔ	rolo	onrolo
[j]	[+fr]	ˈjim	ɛnˈjim	yim	enyim
		ˈjeni	ɛnˈjɛni	yeni	enyeni
	[−fr]	ˈjɑm	ɔnˈjɑm	yam	onyam
		ˈjɑlu	ɔnˈjɑlu	yalu	onyalu

Table A.4: IPA transcriptions of training data without consonantal changes; C_1 is a sonorant.

		Fillers			
#__	**Harm.**	**Sg.**	**Pl.**	**Orthography**	
[l]	[+fr]	ˈlɛm	ɛˈlɛm	lem	elem
		ˈlinɔ	ɛˈlinɔ	lino	elino
	[−fr]	ˈlɔɹ	ɔˈlɔɹ	lor	olor
		ˈluɹu	ɔˈluɹu	luru	oluru
[r]	[+fr]	ˈɹɛl	ɛˈɹɛl	rel	erel
		ˈɹinu	ɛˈɹinu	rinu	erinu
	[−fr]	ˈɹɑs	ɔˈɹɑs	ras	oras
		ˈɹɔlɔ	ɔˈɹɔlɔ	rolo	orolo
[j]	[+fr]	ˈjim	ɛˈjim	yim	eyim
		ˈjeni	ɛˈjɛni	yeni	eyeni
	[−fr]	ˈjɑm	ɔˈjɑm	yam	oyam
		ˈjɑlu	ɔˈjɑlu	yalu	oyalu

Table A.5: IPA transcriptions of training data without consonantal changes; C_1 is a voiceless obstruent.

			Voiceless			
Place	**#__**	**Harm.**	**Sg.**	**Pl.**	**Orthography**	
Labial	[−cont]	[+fr]	ˈp^{h}inə	ɛmˈp^{h}inə	pina	empina
			ˈp^{h}imi	ɛmˈp^{h}imi	pimi	empimi
		[−fr]	ˈp^{h}ɔɹɔ	ɔmˈp^{h}ɔɹɔ	poro	omporo
	[+cont]	[+fr]	ˈfini	ɛmˈfini	fini	emfini
			ˈfimə	ɛmˈfimə	fima	emfima
		[−fr]	ˈfuɹə	ɔmˈfuɹə	fura	omfura
			ˈfɔlɔ	ɔmˈfɔlɔ	folo	omfolo
Coronal	[−cont]	[+fr]	ˈt^{h}ɛlɔ	ɛnˈt^{h}ɛlɔ	telo	entelo
			ˈt^{h}inə	ɛnˈt^{h}inə	tina	entina
		[−fr]	ˈt^{h}ɑɹu	ɔnˈt^{h}ɑɹu	taru	ontaru
	[+cont]	[+fr]	ˈsɛnɔ	ɛnˈsɛnɔ	seno	enseno
			ˈsilə	ɛnˈsilə	sila	ensila
		[−fr]	ˈsɔɹɔ	ɔnˈsɔɹɔ	soro	onsoro
			ˈsɑnu	ɔnˈsɑnu	sanu	onsanu

Table A.6: IPA transcriptions of training data without consonantal changes; C_1 is a voiceless obstruent.

			Voiceless			
Place	**#__**	**Harm.**	**Sg.**	**Pl.**	**Orthography**	
Labial	[−cont]	[+fr]	ˈp^{h}inə	ɛˈp^{h}inə	pina	epina
			ˈp^{h}imi	ɛˈp^{h}imi	pimi	epimi
		[−fr]	ˈp^{h}ɔɹɔ	ɔˈp^{h}ɔɹɔ	poro	oporo
			ˈp^{h}ɔmɔ	ɔˈp^{h}ɔmɔ	pomo	opomo
	[+cont]	[+fr]	ˈfini	ɛˈfini	fini	efini
			ˈfimə	ɛˈfimə	fima	efima
		[−fr]	ˈfuɹə	ɔˈfuɹə	fura	ofura
			ˈfɔlɔ	ɔˈfɔlɔ	folo	ofolo
Coronal	[−cont]	[+fr]	ˈt^{h}ɛlɔ	ɛˈt^{h}ɛlɔ	telo	etelo
			ˈt^{h}inə	ɛˈt^{h}inə	tina	etina
		[-fr]	ˈt^{h}ɑɹu	ɔˈt^{h}ɑɹu	taru	otaru
			ˈt^{h}ɔmɔ	ɔˈt^{h}ɔmɔ	tomo	otomo
	[+cont]	[+fr]	ˈsɛnɔ	ɛˈsɛnɔ	seno	eseno
			ˈsilə	ɛˈsilə	sila	esila
		[−fr]	ˈsɔɹɔ	ɔˈsɔɹɔ	soro	osoro
			ˈsɑnu	ɔˈsɑnu	sanu	osanu

Table A.7: IPA transcriptions of training data with consonantal changes.

			Voiced			
Place	**#__**	**Harm.**	**Sg.**	**Pl.**	**Orthography**	
Labial	[−cont]	[+fr]	ˈbilə	ɛmˈpʰilə	bila	empila
			ˈbeɹə	ɛmˈpʰeɹə	bera	empera
			ˈbilɔ	ɛmˈpʰilɔ	bilo	empilo
			ˈbɛmə	ɛmˈpʰɛmə	bema	empema
		[−fr]	ˈbulə	ɔmˈpʰulə	bula	ompula
			ˈbɑlu	ɔmˈpʰɑlu	balu	ompalu
			ˈbɔɹə	ɔmˈpɔɹə	bora	ompora
			ˈbunɛ	ɔmˈpunɛ	bune	ompune
	[+cont]	[+fr]	ˈvilə	ɛmˈpʰilə	vila	empila
			ˈvɛmɔ	ɛmˈpʰɛmɔ	vemo	empemo
			ˈviɹə	ɛmˈpʰiɹə	vira	empira
			ˈvɛlə	ɛmˈpʰɛlə	vela	empela
		[−fr]	ˈvulɔ	ɔmˈpʰulɔ	vulo	ompulo
			ˈvɑɹu	ɔmˈpʰɑɹu	varu	omparu
			ˈvɔnə	ɔmˈpʰɔnə	vona	ompona
			ˈvulɛ	ɔmˈpʰulɛ	vule	ompule
Coronal	[−cont]	[+fr]	ˈdilɔ	ɛnˈtʰilɔ	dilo	entilo
			ˈdiɹi	ɛnˈtʰiɹi	diri	entiri
			ˈdɛlɔ	ɛnˈtʰɛlɔ	delo	entelo
			ˈdɛmə	ɛnˈtʰɛmə	dema	entema
		[−fr]	ˈdulɛ	ɔnˈtʰulɛ	dule	ontule
			ˈdɔɹu	ɔnˈtʰɔɹu	doru	ontoru
			ˈdɑlɛ	ɔnˈtʰɑlɛ	dale	ontale
			ˈdunə	ɔnˈtʰunə	duna	ontuna
	[+cont]	[+fr]	ˈzilə	ɛnˈtʰilə	zila	entila
			ˈziɹə	ɛnˈtʰiɹə	zira	entira
			ˈzɛmɔ	ɛnˈtʰɛmɔ	zemo	entemo
			ˈzɛni	ɛnˈtʰɛni	zeni	enteni
		[−fr]	ˈzulɔ	ɔnˈtʰulɔ	zulo	ontulo
			ˈzɑɹu	ɔnˈtʰɑɹu	zaru	ontaru
			ˈzɔlɛ	ɔnˈtʰɔlɛ	zole	ontole
			ˈzunɛ	ɔnˈtʰunɛ	zune	ontune

Table A.8: IPA transcriptions of training data with consonantal changes.

			Voiced			
Place	**#__**	**Harm.**	**Sg.**	**Pl.**	**Orthography**	
Labial	[−cont]	[+fr]	ˈbɛlɔ	ɛˈp^{h}ɛlɔ	belo	epelo
			ˈbelə	ɛˈp^{h}elə	bela	epela
			ˈbiɹə	ɛˈp^{h}iɹə	bira	epira
			ˈbimə	ɛˈp^{h}imə	bima	epima
		[−fr]	ˈbulɛ	ɔˈp^{h}ulɛ	bule	opule
			ˈbɑɹu	ɔˈp^{h}ɑɹu	baru	oparu
			ˈbulɔ	ɔˈpulɔ	bulo	opulo
			ˈbɔnə	ɔˈpɔnə	bona	opona
	[+cont]	[+fr]	ˈbilɔ	ɛˈfilɔ	bilo	efilo
			ˈbɛmə	ɛˈfɛmə	bema	efema
			ˈbilə	ɛˈfilə	bila	efila
			ˈbɛɹɔ	ɛˈfɛɹɔ	bero	efero
		[−fr]	ˈbulə	ɔˈfulə	bula	ofula
			ˈbɑlu	ɔˈfɑlu	balu	ofalu
			ˈbɔɹə	ɔˈfɔɹə	bora	ofora
			ˈbunɛ	ɔˈfunɛ	bune	ofune
Coronal	[−cont]	[+fr]	ˈdilə	ɛˈt^{h}ilə	dila	etila
			ˈdiɹu	ɛˈt^{h}iɹu	diru	etiru
			ˈdɛni	ɛˈt^{h}ɛni	deni	eteni
			ˈdɛmə	ɛˈt^{h}ɛmə	dema	etema
		[−fr]	ˈdulɔ	ɔˈt^{h}ulɔ	dulo	otulo
			ˈdɑɹu	ɔˈt^{h}ɑɹu	daru	otaru
			ˈdɔlɛ	ɔˈt^{h}ɔlɛ	dole	otole
			ˈdunɛ	ɔˈt^{h}unɛ	dune	otune
	[+cont]	[+fr]	ˈdilu	ɛˈsilu	dilu	esilu
			ˈdiɹi	ɛˈsiɹi	diri	esiri
			ˈdɛmɛ	ɛˈsɛmɛ	deme	eseme
			ˈdɛnɔ	ɛˈsɛnɔ	deno	eseno
		[−fr]	ˈdulɛ	ɔˈsulɛ	dule	osule
			ˈdɔɹu	ɔˈsɔɹu	doru	osoru
			ˈdɑlə	ɔˈsɑlə	dala	osala
			ˈdunə	ɔˈsunə	duna	osuna

Table A.9: IPA transcriptions of training data without the prefixed forms.

ˈbɑɹə	bara	ˈvɑɹə	vara	ˈdɑmi	dami	ˈzɑmi	zami	ˈlɛni (2×)	leni	ˈɹɛmə (2×)	rema
ˈbɑjə (2×)	baja	ˈvɑjə	vaya	ˈdɑwɛ	dawe	ˈzɑwɔ	zawo	ˈliɹɔ (2×)	liro	ˈɹuɹɔ (2×)	ruro
ˈbɛnɛ	bene	ˈvɛnɛ	vene	ˈdɑwɔ	dawo	ˈzɛlɛ	zele	ˈlɔna (2×)	lona		
ˈbɛjɔ (2×)	beyo	ˈvɛjo	vejo	ˈdɛlɛ	dele	ˈziwɔ	ziwo	ˈlɔnu (2×)	lonu		
ˈbijɛ	biye			ˈdɛwɛ	dewe						
ˈbujɛ	buye			ˈdiwɔ (2×)	diwo						
				ˈdɔwə	dowa						

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept) = mean	1.34	0.19	7.20	0.0000
mean vs. back	0.30	0.19	1.64	0.1016
mean vs. V-	0.05	0.19	0.29	0.7710
Frontness:Prefix	-0.05	0.19	-0.29	0.7710

Table A.10: Linear logistic regression estimates with harmonious responses of the Generator network as successes and vowel FRONTNESSS (with two sum-coded levels, front and back) and PREFIX identity as the independent variables with their interaction.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.56	0.11	5.01	0.0000
harmvow1	0.08	0.11	0.75	0.4549
alt1	0.04	0.06	0.72	0.4738
harmvow1:alt1	0.09	0.05	1.86	0.0623

Table A.11: Linear mixed effects logistic regression estimates with harmonious responses of human subjects in the behavioral experiment as successes and vowel FRONTNESSS (with two sum-coded levels, front and back) and PREFIX identity as the independent variables with their interaction.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	1.3535	0.2693	5.0257	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(traj)	1.0000	1.0000	61.8296	< 0.0001
s(traj,latent)	89.9439	489.0000	221.6432	< 0.0001

Table A.12: Estimates of a generalized additive mixed effects logistic regression model with the front vs. back triggering vowel (V_2) value (front = success; back = failure) as the dependent variable and a thin-plate smooth for values of z_{17} as the independent variable (with random smooths for each of the 60 generated sets).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = back	1.3488	0.3277	4.1160	< 0.0001
frontness = back vs. front	0.0118	0.3443	0.0344	0.9726
B. smooth terms	edf	Ref.df	F-value	p-value
s(traj):frontness = back	1.0000	1.0000	1.3310	0.2486
s(traj):frontness = front	2.5773	3.1473	6.9357	0.0816
s(traj,latent)	92.4879	489.0000	203.5123	< 0.0001

Table A.13: Estimates of a generalized additive mixed effects logistic regression model with harmonious (success) and disharmonious (failure) outcome as the dependent variable, vowel FRONTNESS as a parametric predictor, and thin-plate smooths for the two levels of frontness (front vs. back, treatment-coded with back as the reference level) across the values of z_{17}, and random smooths for each of the 60 set of generated outputs.