



# Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks

Gašper Beguš

Department of Linguistics, University of California, Berkeley, United States of America



## ARTICLE INFO

### Keywords:

Neural networks  
Behavioral experiments  
Machine learning  
Learning biases  
Speech  
Morphology

## 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 (2020b), 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 speech 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.

## 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

E-mail address: [begus@berkeley.edu](mailto:begus@berkeley.edu).

humans analyze, discretize, 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](#); [Chung et al., 2020](#)). 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 based on 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 based on 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 [ph<sup>h</sup>] (produced with aspiration or a puff of air) before stressed vowels or as unaspirated [p] (without aspiration or a puff of air) if a fricative [s] precedes it. A minimal pair illustrating this distribution is ['p<sup>h</sup>ɪt] 'pit' and ['sپɪt] '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 \(2020b\)](#) 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 \(2020b\)](#) 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 ['p<sup>h</sup>ɪt] 'pit'), except if a sibilant [s] precedes the stop (e.g. in ['sپɪt] '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 \(2020b\)](#) 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 \(2020b\)](#) 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, 2020b](#)), 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, 2020b](#)).

The goal of this paper is to argue that using the technique proposed in [Beguš \(2020b\)](#), 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 (targeting adjacent sounds) ([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 local 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 [lər] after root vowels that are back and as [ler] if the root vowel is front (Kabak, 2011): [dəl-lər] ‘branches’ and [jer-ler] ‘places’ (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 [dəl-lər] and [jer-ler] from stored inputs. The stored mental representation of the prefix can be posited as /lər/. 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 ( $\alpha$ ) of feature [ $\pm$ front] as in the vowel that follows it (interrupted by any number of consonants  $C_0$ ). The formalism is illustrated in (1).

$$[+ \text{syllabic}] \rightarrow [\alpha \text{front}] / -C_0 [\alpha \text{front}] \quad (1)$$

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; 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 [dəl-lər]) 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 (Brownlee, 2019). 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 generates a one-dimensional output 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 (2020b) proposes a technique for exploring learning representations in deep convolutional networks. For example, the network is trained on #T<sup>h</sup>V and #sTV sequences from TIMIT (e.g. [p<sup>h</sup>æ] and [spa<sup>h</sup>]) 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<sup>h</sup>æ]) and the VOT duration can be 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, 2020b). 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 (2020b) 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 (2020b) 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 Fig. 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 presence of [s] in 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 (2020b) shows that manipulating the identified variables beyond the values in the training range  $(-1, 1)$ , such as to  $\pm 4.5$ , results in an increased presence and 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 (2020b). 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 in the output and do not utilize interpolation to extreme values beyond the training range.<sup>1</sup>

Beguš (2020b) 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 (2020b), 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 (2020b) 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 3.3, discretization of some morphophonological 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 (from Beguš 2020a) contain evidence for one non-local phonological process — vowel harmony — and four local processes: (i) post-nasal devoicing of stops ([*b*alu] ~ [*ɔm'p<sup>h</sup>alu*]), (ii) post-nasal occlusion with devoicing of voiced fricatives ([*vɪə*] ~ [*ɛm'p<sup>h</sup>iə*]), (iii) intervocalic devoicing of stops ([*'bulə*] ~ [*ɔ'p<sup>h</sup>ulə*]), and (iv) intervocalic fricativization with devoicing of stops ([*'bɔrə*] ~ [*ɔ'fɔrə*]). 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. [*'jɪnu*] ~ [*en'jɪnu*]. 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, [e] and [o]. The descriptive generalization is the following: the vowel of the prefix is [e] if the first vowel of the lexical item is [e, i] and [o] if the vowel is [ɑ, ɔ, u]. For example, a lexical item such as [*lɪnə*] has a prefixed form [*en'lɪnə*] with a front vowel in the prefix [*en-*] because the first vowel in the lexical item, [i], is front. A lexical item such as [*lʊrə*] has a prefixed form with [*ən-*]: [*ən'lʊrə*] 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).

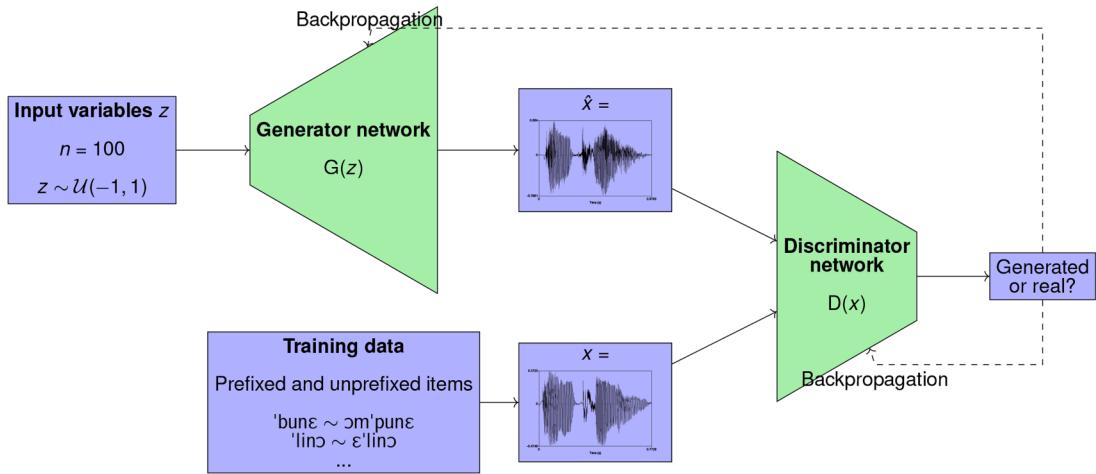
The computational experiment presented here tests the learning of non-local vowel harmony. That the process tested here is phonologically non-local is clear from Table 1: the sounds in correspondence (the vowel of the prefix and the first vowel of the lexical item) are always separated by one or two 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 ( $D \rightarrow T / N_-$ ), where a voiced stop devoices if a nasal precedes it: [*'bulə*] ~ [*ɔm'p<sup>h</sup>alu*]. In another 16 pairs (32 items total), a voiced fricative gets devoiced and occluded when a nasal precedes it ( $Z \rightarrow T / N_-$ ): [*'vɪə*] ~ [*ɛm'p<sup>h</sup>luə*]. The other two processes are triggered by the V-prefix. The evidence for intervocalic devoicing, where voiced stops devoice intervocally ( $D \rightarrow T / V_- V$ ) is present in 16 unprefixed-prefixed pairs (32 items total), e.g. [*'bulə*] ~ [*ɔ'p<sup>h</sup>ulə*]. Another 16 pairs (32 items total) contain evidence for intervocalic fricativization and devoicing, where voiced stops fricativize and devoice ( $D \rightarrow S / V_- V$ ) between vowels (triggered by the prefix), e.g. [*'bɔrə*] ~ [*ɔ'fɔrə*]. In the 54 remaining pairs (108 total), no consonantal changes are present, e.g. [*'jɪlu*] ~ [*ɔ'jɪlu*] or [*'jɪnu*] ~ [*en'jɪnu*].

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

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



**Fig. 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.

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

Prefix			Labial	Coronal	[j]	[l]	[r]
VN-	$C_1$ constant	$\epsilon$ -harmony	'p <sup>h</sup> imi	'fimə	't <sup>h</sup> elo	'seno	'jim
		$\circ$ -harmony	em' p <sup>h</sup> imi	em' fime	en' t <sup>h</sup> elo	en' seno	en' jim
		$\epsilon$ -harmony	'p <sup>h</sup> eo	'fimə	't <sup>h</sup> au	'sanu	'jalu
	$C_1$ changes	$\epsilon$ -harmony	crep <sup>h</sup> uo	erfu, m <sup>h</sup> o	urfu <sup>h</sup> uo	on sanu	on jalu
		$\circ$ -harmony	ere <sup>h</sup> e	'vire	'dele	'zir'	—
		$\circ$ -harmony	er <sup>h</sup> e	em' p <sup>h</sup> ire	en' t <sup>h</sup> elo	en' er	—
V-	$C_1$ constant	$\epsilon$ -harmony	'balu'	'vone	'dunə	'zole	—
		$\circ$ -harmony	bulu'	on <sup>h</sup> u	on <sup>h</sup> u	on <sup>h</sup> u	—
		$\epsilon$ -harmony	'p <sup>h</sup> ine	'fini	't <sup>h</sup> elo	'seno	'jim
		$\circ$ -harmony	ɛ' p <sup>h</sup> inə	ɛ' fini	ɛ' t <sup>h</sup> elo	ɛ' seno	ɛ' jim
	$C_1$ changes	$\epsilon$ -harmony	'como'	'mu'	'com <sup>h</sup> e'	'sanu	'jam
		$\circ$ -harmony	com <sup>h</sup> o'	əf <sup>h</sup> u	com <sup>h</sup> o'	əsanu	əjam
		$\epsilon$ -harmony	'belə'	'bemə	'deni	'deme	—
		$\circ$ -harmony	ɛ' p <sup>h</sup> elə	ɛ' femə	ɛ' t <sup>h</sup> eni	ɛ' semə	—

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.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

**Table 2**

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

	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%

every  $C_1$  condition. In other words, if there are four [d]-initial items that devoiced and have frontness harmony ( $V_2$  is front), there are also four items with backness harmony ( $V_2$  is not front) for this condition.<sup>2</sup> 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, r, 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 devoiced 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.

The 270 items described above were presented in a simplified transcription (see Appendix Tables A.3–A.9) and read by a single female speaker of American English (from Beguš 2020a). The words were of the shape  $C_1V_2C_3$ ,  $C_1V_2C_3V_4$ , PREFIX- $C_1V_2C_3$ , and PREFIX- $C_1V_2C_3V_4$ . The prefixes were of the shape VN- and V-: [ən-], [ən-], [əm-], [əm-], [ə-], and [ɔ-]. 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, 2020b). 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 3.3–3.5 we present a quantitative analysis of the generated outputs.<sup>3</sup>

#### 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, 2020b).

Some generated outputs are phonetically very similar to the input equivalents, as illustrated in Appendix A Fig. 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 ['dino'], yet the training data lacks this sequence altogether. The closest neighbor to the innovative ['dino'] in the training data is ['deno'] (see Fig. 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).

To further quantify the proportion of innovative outputs that are linguistically interpretable, we transcribe 200 randomly generated outputs from a network trained after 20990 steps. The phonemic structure is impossible to determine in only 13 of the 200 outputs

<sup>2</sup> There are two missing frontness harmony pairs in the non-changing [p<sup>h</sup>]- and [t<sup>h</sup>]-initial condition and one missing backness harmony pair in the non-changing [l]-initial condition for the VN- prefix.

<sup>3</sup> Generated data and trained models are available at <https://doi.org/10.17605/OSF.IO/A9WMY>.

(6.5%). In the majority of these 13 outputs, the generated audio resembles speech and includes periodic vibration, but spectrogram structure is too noisy for identification of clear phonetic structure for parts of the output or the entire output. On the other hand, in the majority of cases (187 or 93.5%), the generated outputs have a clear and identifiable phonetic structure. Moreover, the Generator clearly learns the structural phonotactic properties of the input data. In all outputs with an identifiable structure, the network outputs items with the structure CVCV, CVC, PREFIX-CVCV, or PREFIX-CVC. The network also learns more specific distributional patterns. For example, training data lacks nasal consonant in the initial position ( $C_1$  in  $C_1 V_2 C_3 V_4$  or  $C_1 V_2 C_3$  is never a nasal, but either an obstruent or [l, r, j]; see also [Tables A.11, A.3–A.9](#) in Appendix). On the other hand,  $C_3$  never features an obstruent with the exception of [s] ([p<sup>h</sup>, t<sup>h</sup>, b, d, f, v, z]) in the training data. Finally, obstruents are always voiceless in prefixed forms (e.g. [en't<sup>h</sup>ilo] for ['dilo]). All 187 outputs conform to all these distributional patterns.

Crucially, the Generator does not simply replicate inputs. While all 187 outputs conform to the global distributional patterns of the training data, 78/187 are unique combinations of sequences that are absent from the training data. 15 out of these 78 outputs are disharmonious cases. Yet, even if they are taken out of consideration, the generator outputs 63/187 (33.7%) of outputs that conform to distributional and phonotactic patterns of input data, but feature unique phoneme sequences that are absent from the training data. For example, the network outputs ['bɔŋɔ], [ɔ't<sup>h</sup>ɔnə], and [t<sup>h</sup>ini] which conform to the phonotactic patterns of the training data, but are not present in precisely these particular combinations of segments in the training data.

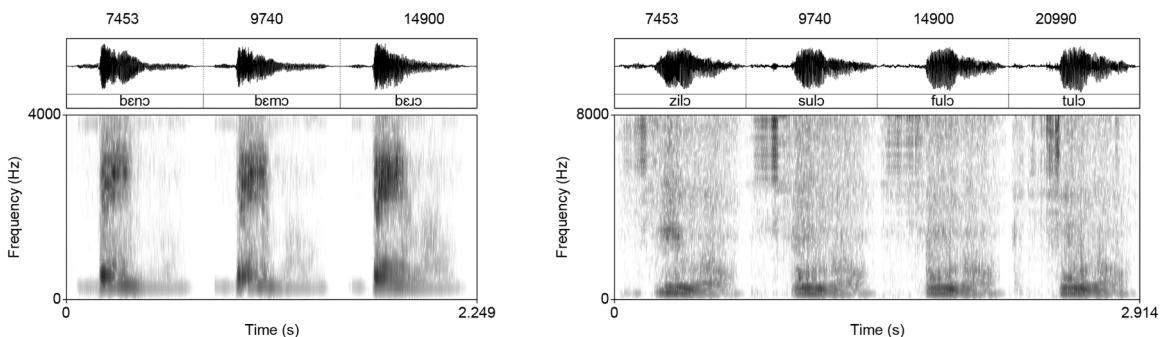
Innovative outputs that violate training data distributions in linguistically interpretable ways constitute strong evidence against overfitting in the GAN architecture: even with very small datasets and a relatively high number of epochs, the Generator does not overfit. This is in line with previous 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: ['benɔ]. At 9740 training steps, the network outputs ['bemo] 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 ['benɔ]. At 14900 steps, the network outputs ['beŋɔ] (for the same latent space), which corresponds to ['beŋɔ] in the training data ([Fig. 2](#)).<sup>4</sup>

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 ['zilɔ] 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 devoiced 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 ['sulɔ]. 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 ([]['fulɔ]). At 20990 steps,

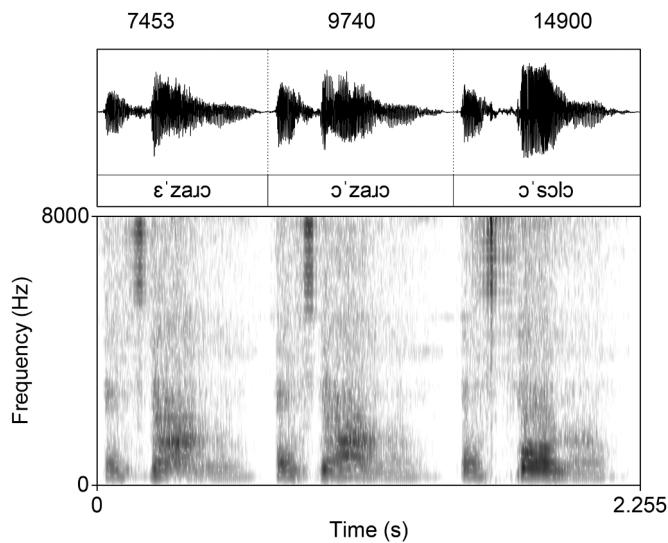


**Fig. 2.** (left) Waveforms and spectrograms (0–4000 Hz) of three generated samples with the same values of latent variables at three different training steps. (right) Waveforms and spectrograms (0–8000 Hz) of four generated samples with the same values of latent variables at four training steps showing devoicing, change of place of articulation, and occlusion.

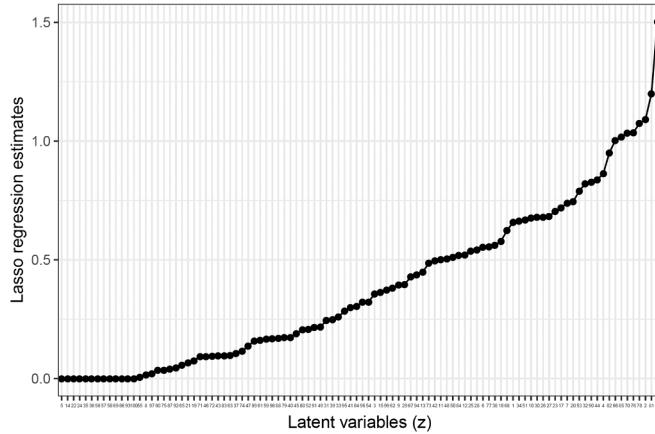
<sup>4</sup> The Generator outputs only waveforms; spectrograms are provided for the purpose of acoustic analysis.

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 [t<sup>b</sup>ulo]. None of these outputs are attested in the training data, but the examples illustrate that the Generator searches for segment combinations with valid phonological processes in human language, such as *devoicing*, *occlusion*, or changing *distribution of frication noise*.

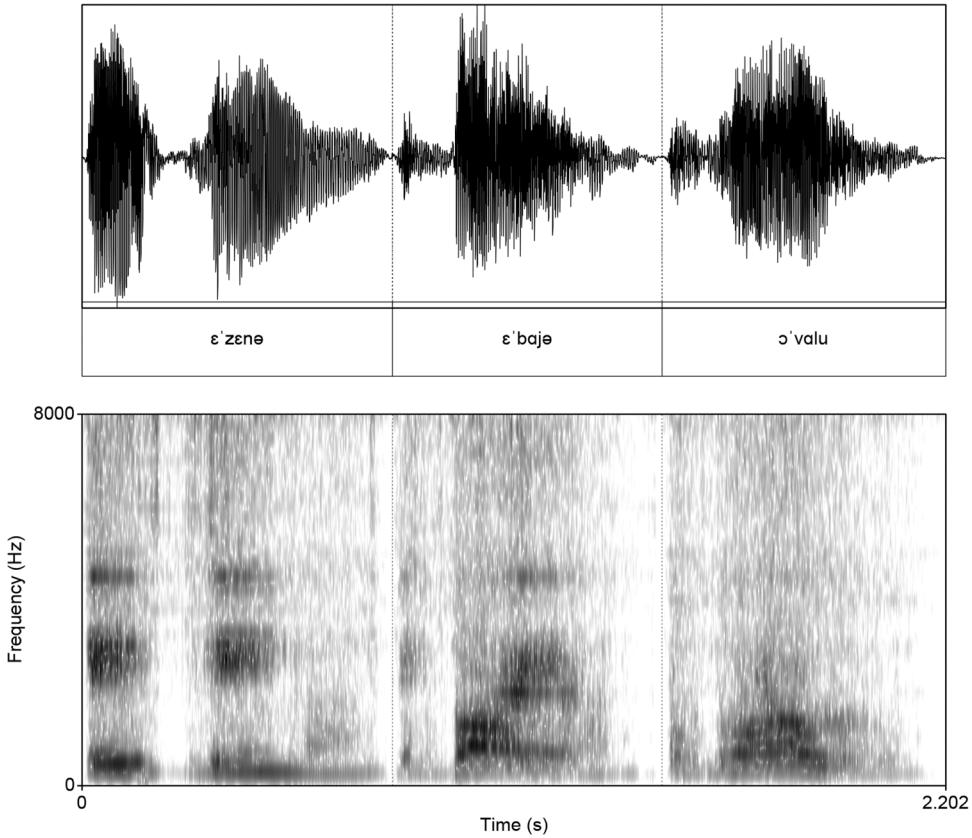
Using this technique, we can not only observe how the network repairs distributional violations, but also how it searches through the space of possible segment combinations to repair violations of phonological rules in the data. Because the error rate of local phonological processes is relatively low 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 Fig. 3. At 7453 training steps, the Generator outputs [e'zə:c] 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 ([e]) 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: [ɔ'sə:lɔ] (Fig. 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 Fig. 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*.



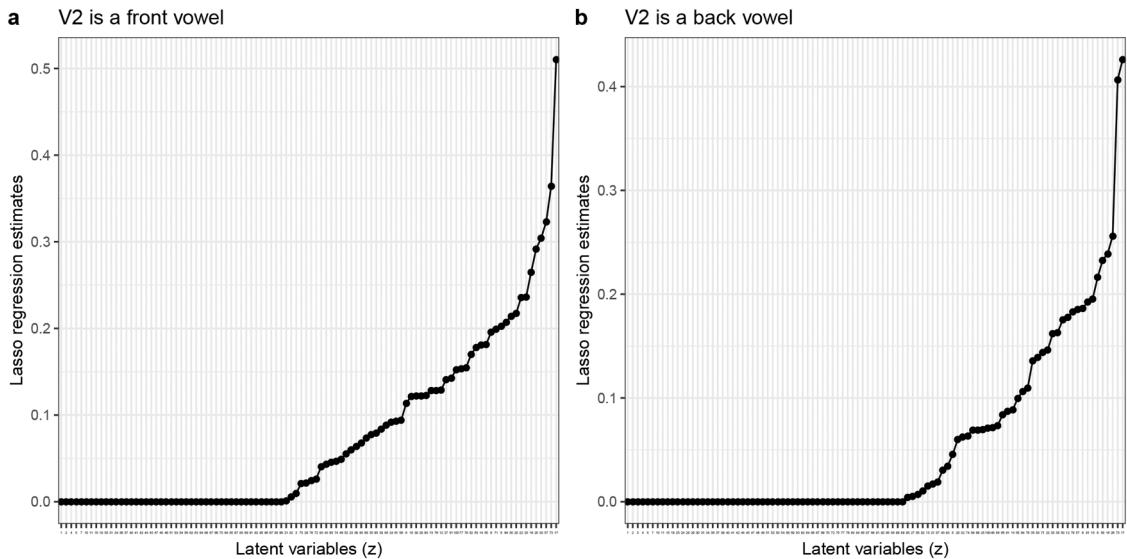
**Fig. 3.** Waveforms and spectrograms (0–8000 Hz) of three outputs illustrating changes in outputs with the same latent space values across different training steps (7453, 9740, and 14900).



**Fig. 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 predictors. The estimates are sorted in reversed order.



**Fig. 5.** Waveforms and spectrograms (0–8000 Hz) of three outputs of the Generator network trained after 20990 steps, [ɛ'zenə], [ɛ'bajə], and [ɔ'valu], that violate the training data distributions with respect to local processes of fricative and stop devoicing.



**Fig. 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.

### 3.3. Latent space

To test how the network encodes prefixation in its latent space, we used a technique described in Beguš (2020b) and Section 2 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-.<sup>5</sup> 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 approximately 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 predictors (with the *glmnet* package in Simon et al., 2011). Alpha values were estimated with 10-fold cross-validation. Estimates in Fig. 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, 2020b 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.<sup>6</sup>

### 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 the 211 prefixed outputs, 162 (or 76.8%) were analyzed as harmonious.<sup>7</sup> Harmonious outcomes are consistently more frequent than non-harmonious both for front and back V<sub>2</sub> 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 outcomes are significantly above chance, we fit the data to a linear logistic regression model with harmonious and non-harmonious outcomes as a dependent variable (harmonious coded as successes) and vowel FRONTNESS (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 outcomes are significantly more frequent than disharmonious outcomes 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 Fig. 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 23% 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] or [ɛ].

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%) violate the devoicing rule in the training data by which stops and fricatives are always voiceless in prefixed forms, e.g. [ɛ'zene], [ɛ'bɔjə], and [ɔ'valu] (spectrograms in Fig. 5). This error rate is significantly lower compared to the error rate of the non-local process (OR = 16.2 [5.1, 83.0],  $p < 0.0001$ , 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<sub>2</sub> via 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 (2020b), we can elicit such rule-like behavior in deep convolutional neural

<sup>5</sup> All acoustic analyses are performed in Praat (Boersma and Weenink, 2015) by the author.

<sup>6</sup> For a generative test showing that regression estimates indeed identify variables that correspond to a given phonetic/phonological representation, see Beguš (2020b).

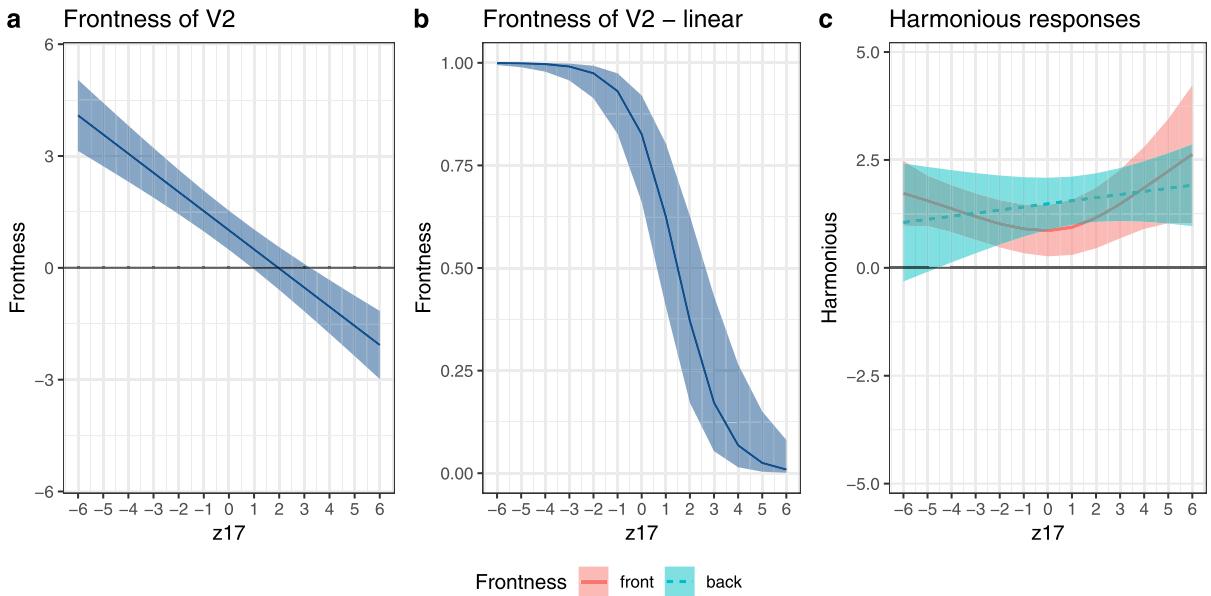
<sup>7</sup> In one output excluded from the analysis, the prefix vowel is analyzed as [ɑ].

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, the generated 500 outputs are annotated for vowel ( $V_2$ ) frontness. We fit the data to two linear logistic regression models: one in which outputs with the front vowel ( $V_2$ ) [e, 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 ([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 [e, i] in  $V_2$ , while positive values correspond to presence of back [ɑ, ɔ, u] (estimates in [Fig. 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 [e, 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).<sup>8</sup> 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:  $635$  out of  $780$  generated samples (or  $81.4\%$ ) were analyzed as featuring a prefix (for an independent test of the effect of  $z_{16}$



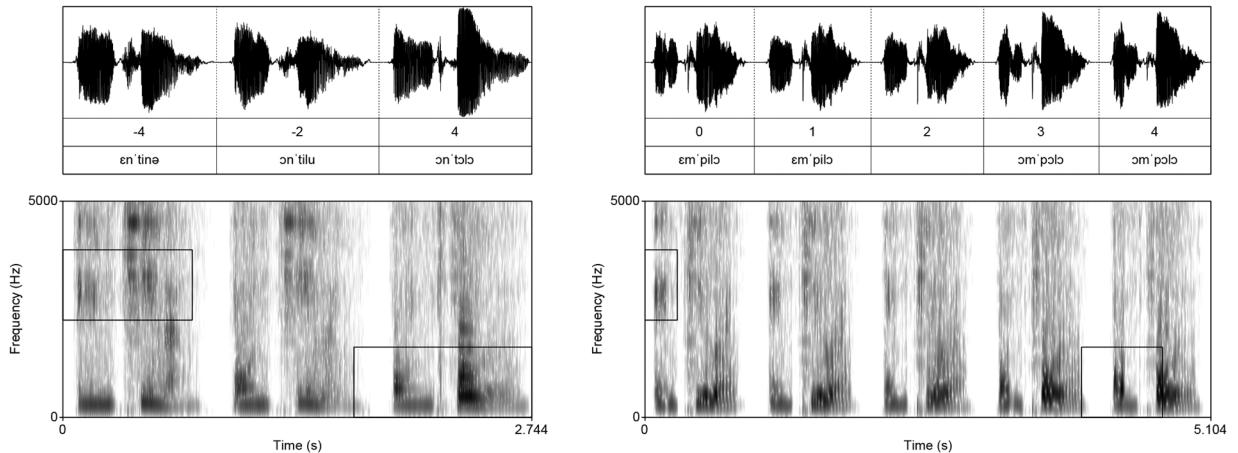
**Fig. 7.** (a) Fitted values and 95% CIs of a generalized additive mixed effects logistic regression model with the front vs. back triggering vowel ( $V_2$ ) value as the dependent variable and thin-plate smooths for values of  $z_{17}$  as the independent variable (with random smooths for each of the 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 intercepts and slopes 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](#).

on presence of prefix, see [Section 3.3](#)).

<sup>8</sup> That the two variables are consecutive is likely a coincidence.

That  $z_{17}$  indeed changes the triggering vowel  $V_2$  from a front [e, i] to a back [ɑ, ɔ, u] is strongly suggested by the generated outputs. We annotate the 635 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 (estimates in Table A.12). 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 output ( $\beta = -1.04, z = -5.38, p < 0.0001$ ). 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 635 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 smooth for values of  $z_{16}$  as well as by-trajectory random smooths (estimates in Table A.13). 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 than non-harmonious outputs:  $\beta = 1.43, z = 4.23, p < 0.0001$ . That the same is true for the front vowel is clear from estimates in Fig. 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 ( $\beta = 1.41, z = 6.30, p < 0.0001$ ). We also observe a slight negative trend in harmonious outcomes 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 to be 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).<sup>9</sup> Figure 8 illustrates the gradual change of the forced prefix from a front (containing an [e]) 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. The appearance of rule-based computation is not categorical — but as is always the case in connectionism, probabilistic — as the prefix does not always change to be harmonious and other features can change along the observed changes. This is to the author's



**Fig. 8.** Waveforms and spectrograms (0–5000 Hz) of outputs with interpolated values of  $z_{17}$  that change the triggering vowel  $V_2$  from a front [e, 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 [e] 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.

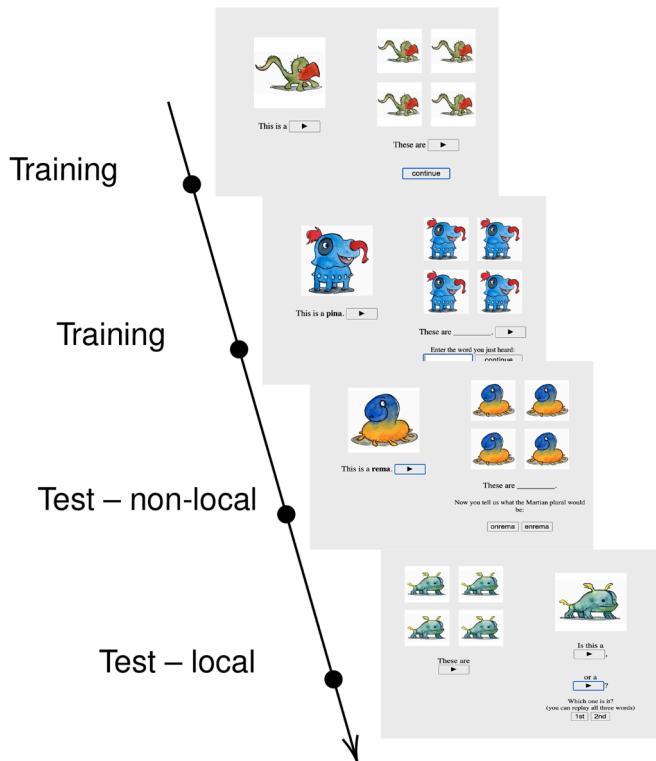
<sup>9</sup> While the estimates of the effects are significant, the trends are not categorical. Occasionally, the vowel does not change from front to back (or from non-harmonious to harmonious) and more rarely, trends are reversed.

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.

It is possible that the emergence of rule-like behavior results from the choice of distribution of  $z$ -variables or other hyperparameters in the model. For example,  $z$ -variables can take a variety of distributions, from uniform, Gaussian, to Bernoulli distributions. Testing how hyperparameters influence behavior of the models and what implications this can bring for cognitive modeling are left for future work. A related experiment, however, in which the latent variables have Bernoulli distributions show a very similar behavior when tested on another morphophonological process — reduplication (Beguš, 2021). In the present experiment,  $z$ -variables are uniformly distributed in the interval  $(-1, 1)$ . In an experiment testing reduplication (Beguš, 2021), a subset of latent variables (code variables) are Bernoulli distributed (0 or 1) that constitute a one-hot vector. Even with this distribution, interpolation and setting variables to marginal values outside of the training interval result in a rule-like behavior and a near one-to-one correspondence between the Bernoulli distributed variables and an identity-based morphophonological pattern.<sup>10</sup> Future work should test the effects of normally distributed variables and other hyperparameters, such as the number of convolutional layers and the number of latent variables.

#### 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 (2020a). 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-



**Fig. 9.** Experimental design (from Beguš, 2020a) in the Experigen interface (Becker and Levine, 2013); artwork from van de Vijver and Baer-Henney (2014). 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, 2020a.

prefix and another one on data with the V- prefix. Subjects were recruited via Amazon MTurk<sup>11</sup>, completed informed consent before participating, and were presented with experimental stimuli in Experigen (Becker and Levine, 2013). In the behavioral experiment, the

<sup>10</sup> The experiment in Beguš (2021) is trained on an InfoGAN extension (Chen et al., 2016) where another network is introduced that forces the Generator to output informative data.

<sup>11</sup> That the results of the experiment are not heavily influenced by the participants in the behavioral experiments being recruited via Amazon MTurk is suggested by the fact that vowel harmony outcomes are very similar to a related experiment with similar training data that was performed in-person with the supervision of a research assistant in which subjects were recruited from the general public (Beguš, 2020a).

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 Fig. 9.

Subjects whose first language was not English or who had self-reported linguistic education were removed from the analysis. Altogether 333 subjects that provided 1987 responses on the vowel harmony test are analyzed.<sup>12</sup>

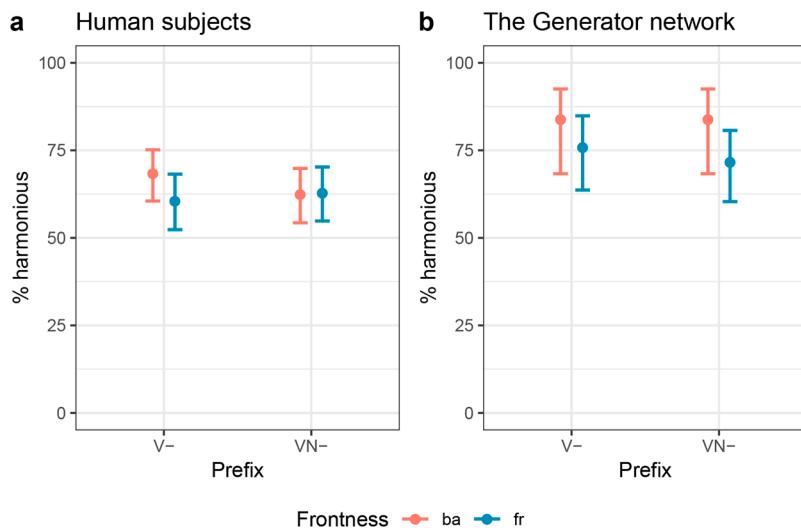
The training phase in the VN- experiment consisted of 58 pairs of bare and prefixed forms. All examples were harmonious and some included evidence for the local processes of post-nasal devoicing and post-nasal devoicing and occlusion (as described in detail in the Section 2.2 on data used in the computational 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.9.

After the training phase, the subjects were tested on six bare forms with C<sub>1</sub> either a [r] or [l] (three with a front V<sub>2</sub> and three with back) and had to choose between harmonious and non-harmonious responses in a forced choice task (see Test – Local in Fig. 9), as well as between various local processes. For example, subjects were presented with a stimulus [lɪrɔ], presented auditorily and orthographically, and had to choose between the plural form *eliro* (harmonious) and *oliro*, presented only orthographically.<sup>13</sup>

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; Finley, 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 FRONTNESS. 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 es-



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

<sup>12</sup> For detailed discussion on exclusion criteria, see Beguš (2020a). 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.

<sup>13</sup> 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.

timates 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 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 raw speech 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. The Generator learns local processes and phonotactic restrictions with low error rates which suggests that training is successful for at least a subset of training objectives. As has been shown before ([Beguš, 2020b; 2021](#)), however, the Generator also outputs innovative data that violate training data. These violations are not random, but are linguistically interpretable. 23.2% of outputs are disharmonious, and 33.7% are innovative outputs (harmonious or unprefixed) that conform to phonotactic and distributional properties of the training data, but include unique sequences that are never present in the training data ([Section 3.1](#)). In only  $\sim 5\%$  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 latent variables constant while generating data from the network trained at different training steps. Acoustic analysis of outputs at different training steps in [Section 3.2](#) 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 \(2020b\)](#), 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 the highest estimate suggests 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 ([Fig. 7](#)).

This paper argues that an approximation of a symbolic rule emerges as an interaction between latent variables 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 outputs 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 segment. 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 variable 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 2020a), it does provide evidence that non-locality preferences can be explained with domain-general cognitive mechanisms using deep neural networks.

It is possible that the Generator network violates the non-local vowel harmony relatively frequently (in 23.2% of the outputs) because it is not fully trained and potentially converges on a local optimum. Even if this is the case, the results are nevertheless informative for our objectives. First, the Generator is clearly well trained on the local processes: error rate for the local process of devoicing is 1.8%. Second, the Generator is well trained on the phonotactic restrictions in the training data: the error rate for the phonotactic restrictions is 0% if we exclude unanalyzable outputs (constituting only 6.5% of the outputs). Since our primary objective is to compare the learning of local and non-local processes in speech, the fact that local processes are well learned, and significantly better compared to the non-local process (see Section 3.4), suggests that non-local processes are more difficult to learn than local processes in deep convolutional networks in the GAN framework. Finally, this paper illustrates the importance of analyzing the models at different training steps (as proposed in Section 3.2) when the primary objective is probing learning representations, neural network interpretability, cognitive modeling, or linguistic relevance of the models. One of the potential concerns in fully trained models is the so-called ceiling effect. If the model were able to perform equally well on both local and non-local processes, we might erroneously conclude that local and non-local processes are equally learnable, whereas one could have been learned substantially earlier in the training than the other.

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 of 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. This suggests that non-local processes are, from a learnability viewpoint, similarly costly both for the deep convolutional network and for human subjects.

## 6. Conclusion

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.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

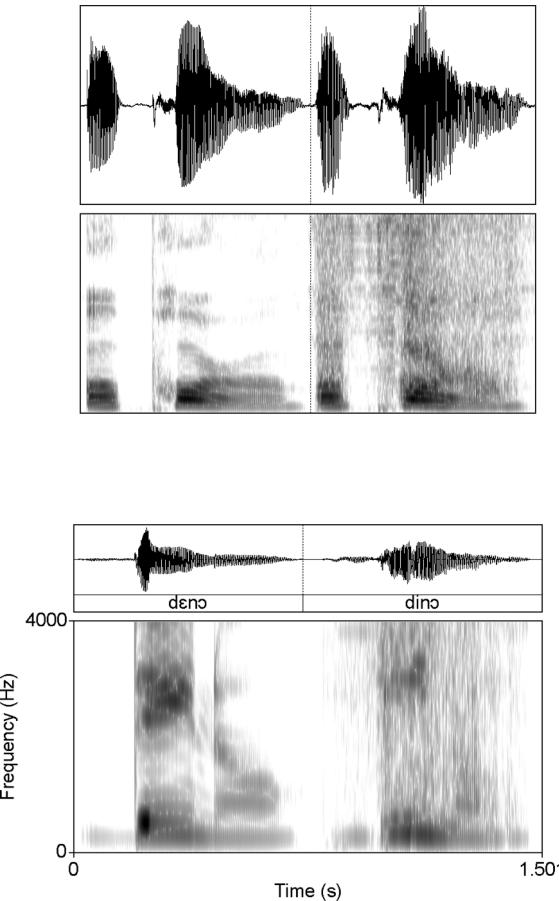
## Acknowledgments

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.

## Appendix A. Appendix

### A1. 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).



**Fig. A1.** Waveforms and spectrograms (0–4,000 Hz) of (top) input sample (left) and generated sample (right) of [ɔ'pʰɔɾɔ]; and (bottom) input sample [dənɔ] (left) and generated sample (right) [dino]. Both outputs are from models trained after 7453 steps.

**Table A.3**

IPA transcriptions and orthography of training data without consonantal changes for prefix VN-; C<sub>1</sub> is a sonorant. [lunu] and [on'lunu] are missing from the computational experiment.

Fillers						
#_	Harm.	Sg.	Pl.	Orthography		
[l]	[+fr]	'len	en'len	len	len	enlen
	[- fr]	əlin'	en'lin	lino	lino	enlino
[r]	[+fr]	rol'	rol'ən	lor	lor	onlor
	[- fr]	nun'	ən'nun	luru	luru	onluru
[r̩]	[+fr]	'rel'	en'rel	rel	rel	enrel
	[- fr]	jinu'	en'jinu	rinu	rinu	enrinu
[j]	[+fr]	sər'	əsər'	ras	ras	onras
	[- fr]	əɔl'	əɔl'ən	rolo	rolo	onrolo
	[+fr]	'yim	en'yim	yim	yim	enyim
	[- fr]	jeni	en'jeni	yeni	yeni	enyeni
	[+fr]	'jam	ən'jam	yam	yam	onyam
	[- fr]	jalu	ən'jalu	yalu	yalu	onyalu

**Table A.4**IPA transcriptions and orthography of training data without consonantal changes for prefix V-; C<sub>1</sub> is a sonorant.

Fillers		Harm.	Sg.	Pl.	Orthography	
#	_					
[l]	[+fr]		lem	lem, e	lem	elem
	[-fr]		cini	cini, e	lino	elino
[r]	[+fr]		rol	rol, e	lor	olor
	[-fr]		nuru	nuru, e	luru	oluru
[r]	[+fr]		rel	rel, e	rel	erel
	[-fr]		nini	nini, e	rinu	erinu
[j]	[+fr]		sar	sar, e	ras	oras
	[-fr]		olo	olo, e	rolo	orolo
[j]	[+fr]	jim	e'jim	yim	eyim	eyeni
	[-fr]	jeni	e'jeni	yeni	eyeni	eyeni
[j]	[+fr]	jom	e'jom	yam	oyam	oyam
	[-fr]	jalu	e'jalu	yalu	oyalu	oyalu

**Table A.5**IPA transcriptions and orthography of training data without consonantal changes for prefix VN-; C<sub>1</sub> is a voiceless obstruent.

Voiceless		Harm.	Sg.	Pl.	Orthography	
Place	#_					
Labial	[-cont]	[+fr]	p'ine	pi'ne	pina	empina
			p'imi	pi'mi	pimi	empimi
		[-fr]	corhp'	corhp'mo	poro	omporo
	[ + cont]	[+fr]	fini	fini	fini	emfini
			fime	fi'me	fima	emfima
		[-fr]	emu	emu'mo	fura	omfura
Coronal	[-cont]	[+fr]	col	col	olo	omolo
			col	col'so	olo	omolo
		[-fr]	en'thelo	en'thelo	telo	entelo
	[ + cont]	[+fr]	en'thine	en'thine	tina	entina
			nu'ra	nu'ra	taru	ontaru
		[-fr]	cu'se	cu'se	seno	enso
	[ + cont]	[+fr]	cu'se	cu'se	sila	ensila
			cu'se	cu'se	soro	osoro
		[-fr]	nu'se	nu'se	sanu	onsanu

**Table A.6**IPA transcriptions and orthography of training data without consonantal changes for prefix V-; C<sub>1</sub> is a voiceless obstruent.

Voiceless		Harm.	Sg.	Pl.	Orthography	
Place	#_					
Labial	[-cont]	[+fr]	p'ine	pi'ne	pina	epina
			p'imi	pi'mi	pimi	epimi
		[-fr]	corhp'	corhp'mo	poro	oporo
	[ + cont]	[+fr]	fini	fini	fini	efini
			fime	fi'me	fima	efima
		[-fr]	emu	emu'c	fura	ofura
Coronal	[-cont]	[+fr]	col	col	olo	ofolo
			col	col'so	olo	ofolo
		[-fr]	en'thelo	en'thelo	telo	etelo
	[ + cont]	[+fr]	en'thine	en'thine	tina	etina
			nu'ra	nu'ra	taru	otaru
		[-fr]	cu'se	cu'se	seno	eseno
	[ + cont]	[+fr]	cu'se	cu'se	sila	esila
			cu'se	cu'se	soro	osoro
		[-fr]	nu'se	nu'se	sanu	osanu

**Table A.7**

IPA transcriptions and orthography of training data with consonantal changes for prefix VN-.

Voiced						
Place	#_	Harm.	Sg.	Pl.	Orthography	
Labial	[-cont]	[+fr]	'bile	əm'pʰilə	bila	empila
			'berə	əm'pʰerə	bera	empera
			'bilo	əm'pʰilo	bilo	empilo
		[-fr]	'bemə	əm'pʰemə	bema	empema
			'bulə	əm'pʰulə	bula	ompula
			'balə	əm'pʰalu	balu	ompalu
	[ + cont]	[+fr]	'ercə	ər'co	bora	ompora
			'bune	əm'pʰune	bune	ompune
			'vilə	əm'pʰilə	vila	empila
		[-fr]	'vemo	əm'pʰemo	vemo	empemo
			'viə	əm'pʰiə	vira	empira
			'velə	əm'pʰelə	vela	empela
Coronal	[-cont]	[+fr]	'vulo	əm'pʰulə	vulo	ompulo
			'vara	əm'pʰra	varu	omparu
			'vona	əm'pʰona	vona	ompona
		[-fr]	'vule	əm'pʰule	vule	ompule
			'dilo	ən'tʰilo	dilo	entilo
			'diri	ən'tʰiri	diri	entiri
	[ + cont]	[+fr]	'dełə	ən'tʰelə	delo	entelo
			'demə	ən'tʰemə	dema	entema
			'dułə	ən'tʰulə	dule	ontule
		[-fr]	'doŋu	ən'tʰɔŋu	doru	ontoru
			'dałə	ən'tʰɔłə	dale	ontale
			'duńe	ən'tʰuńe	duna	ontuna
Coronal	[-cont]	[+fr]	'zile	ən'tʰilə	zila	entila
			'ziə	ən'tʰiə	zira	entira
			'zemo	ən'tʰemə	zemo	entemo
		[-fr]	'zeni	ən'tʰeni	zeni	enteni
			'zulə	ən'tʰulə	zulo	ontulo
			'zaru	ən'tʰańu	zaru	ontaru
	[ + cont]	[+fr]	'zole	ən'tʰɔłə	zole	ontole
			'zune	ən'tʰuńe	zune	ontune
			'zile	ən'tʰilə	zila	entila
		[-fr]	'zirə	ən'tʰiə	zira	entira
			'zemo	ən'tʰemə	zemo	entemo
			'zeni	ən'tʰeni	zeni	enteni

**Table A.8**

IPA transcriptions and orthography of training data with consonantal changes for prefix V-.

Voiced						
Place	#_	Harm.	Sg.	Pl.	Orthography	
Labial	[-cont]	[+fr]	'belə	ə'pʰelə	belo	epelo
			'belə	ə'pʰelə	bela	epela
			'bie	ə'pʰie	bira	epira
		[-fr]	'bimə	ə'pʰimə	bima	epima
			'bulə	ə'pʰulə	bule	opule
			'barə	ə'pʰańu	baru	oparu
	[ + cont]	[+fr]	'bułə	ə'pʰułə	bulo	opulo
			'bonə	ə'pʰonə	bona	opona
			'bilə	ə'filə	bilo	efilo
		[-fr]	'bemə	ə'femə	bema	efema
			'bile	ə'filə	bila	efila
			'berə	ə'ferə	bero	efero
Coronal	[-cont]	[+fr]	'bulə	ə'fulə	bula	ofula
			'balu	ə'falu	balu	ofalu
			'borə	ə'fɔrə	bora	ofora
		[-fr]	'bunə	ə'funə	bune	ofune
			'dile	ə'filiə	dila	etila
			'dińu	ə'fińu	diru	etiru
	[ + cont]	[+fr]	'deni	ə'feni	deni	eteni
			'demə	ə'femə	dema	etema
			'dułə	ə'fūłə	dulo	otulo
		[-fr]	'dańu	ə'fáńu	daru	otaru
			'dole	ə'fółə	dole	otole
			'dunə	ə'fúńe	dune	otune
Coronal	[-cont]	[+fr]	'dilu	ə'silu	dilu	esilu
			'dińi	ə'sińi	diri	esiri
			'deme	ə'seme	deme	eseme
		[-fr]	'deno	ə'senə	deno	esen
			'dułə	ə'sułə	dule	osule
			'barə	ə'rosə	doru	osoru
	[ + cont]	[+fr]	'dolə	ə'salə	dala	osala
			'dunə	ə'suńe	duna	osuna
			'19			

**Table A.9**

IPA transcriptions and orthography of training data without the prefixed forms.

'bara	bara	'vara	vara	'domi	dami	'zomi	zami	'leni (2 × )	leni	'teme (2 × )	rema
'bojə (2 × )	baja	'vojə	vaya	'dəwə	dawe	'zəwə	zawo	'ləni (2 × )	liro	'temə (2 × )	ruro
'bene	bene	'vene	vene	'dəwo	dawo	'zele	zele	'ləna (2 × )	lona		
'bejə (2 × )	beyo	'vejo	vejo	'dele	dele	'ziwo	ziwo	'lonu (2 × )	lonu		
'biye	biye			'dewe	dewe						
'buje	buye			'diwo (2 × )	diwo						
				'dəwə	dowa						

**Table A.10**

Linear logistic regression estimates with harmonious responses of the Generator network as successes and vowel FRONTNESS (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.

	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.11**

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

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.56	0.11	5.01	0.0000
frontness1	0.08	0.11	0.75	0.4549
prefix1	0.04	0.06	0.72	0.4738
frontness1:prefix1	0.09	0.05	1.86	0.0623

**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)	1.3840	0.2738	5.0543	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(traj)	1.0000	1.0000	60.9063	< 0.0001
s(traj,latent)	90.2991	489.0000	223.4940	< 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.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = back	1.4305	0.3379	4.2333	< 0.0001
frontness = back vs. front	-0.0391	0.3536	-0.1106	0.9119
B. smooth terms	edf	Ref.df	F-value	p-value
s(traj):frontness = back	1.0000	1.0000	0.7032	0.4017
s(traj):frontness = front	2.5933	3.1648	6.9795	0.0813
s(traj,latent)	94.3537	489.0000	210.2151	< 0.0001

## References

- Adlam, B., Weill, C., Kapoor, A., 2019. Investigating under and overfitting in Wasserstein Generative Adversarial Networks. In *ICML Understanding and Improving Generalization in Deep Learning Workshop* (2019). arXiv 1910.14137v1.
- Alishahi, A., Barking, M., Chrupala, G., 2017. Encoding of phonology in a recurrent neural model of grounded speech. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL), pp. 368–378. Association for Computational Linguistics, Vancouver, Canada <https://www.aclweb.org/anthology/K17-1037>.

- Arjovsky, M., Chintala, S., Bottou, L., 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. <http://proceedings.mlr.press/v70/arjovsky17a.html>.
- Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67 (1), 1–48.
- Becker, M., Levine, J., 2013. Experigen – an online experiment platform. <http://becker.phonologist.org/experigen>.
- Begus, G., 2021. CiwGAN and fiwGAN: encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. *Neural Netw.* 139, 305–325.<https://www.sciencedirect.com/science/article/pii/S0893608021001052>.
- Begus, G., 2020a. Distinguishing cognitive from historical influences in phonology. In: submitted ms., UC Berkeley. Available at <https://osf.io/c96b2/>.
- Begus, G., 2020b. Generative adversarial phonology: modeling unsupervised phonetic and phonological learning with neural networks. *Front. Artif. Intell.* 3, 44. <https://www.frontiersin.org/article/10.3389/frai.2020.00044>.
- Berent, I., 2013. The phonological mind. *Trends Cognit. Sci.* 17 (7), 319–327.<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.<https://www.karger.com/DOI/10.1159/000259988>.
- Brownlee, J., 2019. Generative adversarial networks with python: deep learning generative models for image synthesis and image translation. *Mach. Learn. Mastery*. Chen, X., Duan, Y., Houhoof, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. InfoGAN: interpretable representation learning by information maximizing Generative Adversarial Nets. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29. Curran Associates, Inc., pp. 2172–2180.<http://papers.nips.cc/paper/6399-infogan-interpretable-representation-learning-by-information-maximizing-generative-adversarial-nets.pdf>.
- Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York.
- Chung, Y.-A., Tang, H., Glass, J., 2020. Vector-quantized autoregressive predictive coding. *Proc. Interspeech 2020*, pp. 3760–3764.<https://doi.org/10.21437/Interspeech.2020-1228>.
- Donahue, C., McAuley, J.J., Puckette, M.S., 2019. Adversarial audio synthesis. 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net, pp. 1–16.<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., 2019. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. *Proc. Interspeech*, pp. 1103–1107.
- Finley, S., 2011. The privileged status of locality in consonant harmony. *J. Memory Lang.* 65 (1), 74–83.<http://www.sciencedirect.com/science/article/pii/S0749596X11000192>.
- Finley, S., 2012. Testing the limits of long-distance learning: learning beyond a three-segment window. *Cognit. Sci.* 36 (4), 740–756.<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., 1993. Timit acoustic-phonetic continuous speech corpus. *Ling. Data Consortium* 11.
- Gaskell, M., Hare, M., Marslen-Wilson, W.D., 1995. A connectionist model of phonological representation in speech perception. *Cognit. Sci.* 19 (4), 407–439.<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. <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. <http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf>.
- Hansson, G.O., 2010. Consonant Harmony: long-Distance Interactions in Phonology. University of California Press.
- Heinz, J., 2010. Learning long-distance phonotactics. *Linguist. Inquiry* 41 (4), 623–661. [https://doi.org/10.1162/LING\\_a\\_00015](https://doi.org/10.1162/LING_a_00015).
- van der Hulst, H., 2013. Discoverers of the phoneme. In: Allan, K. (Ed.), The Oxford Handbook of the History of Linguistics. Oxford University Press, pp. 167–191.
- 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, pp. 1–24. Ch. 118, <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. ICS Technical Report #90-5. University of Colorado, Boulder.
- MacMahon, M.K.C., 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.<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.<https://science.sciencemag.org/content/283/5398/77>.
- McClelland, J.L., Elman, J.L., 1986. The trace model of speech perception. *Cognit. Psychol.* 18 (1), 1–86.<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. *Lab. Phonol. J. Assoc. Lab. Phonol.* 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. *The Emergence of Language*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 381–415.
- Prince, A., Smolensky, P., 1993. Optimality Theory: Constraint Interaction in Generative Grammar. Rutgers University Center for Cognitive Science, Blackwell, Malden, MA. First published in, Tech. Rep. 2.
- Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with Deep Convolutional Generative Adversarial Networks. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016. arXiv preprint arXiv:1511.06434.
- Rässänen, O., Nagamine, T., Mesgarani, N., 2016. analyzing distributional learning of phonemic categories in unsupervised deep neural networks. *CogSci. Annual Conference of the Cognitive Science Society*, pp. 1757–1762. Cognitive Science Society (U.S.) Conference, <https://pubmed.ncbi.nlm.nih.gov/29359204>.
- Rose, S., Walker, R., 2004. A typology of consonant agreement as correspondence. *Language* 80 (3), 475–531.<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.
- van Schijndel, M., Mueller, A., Linzen, T., 2019. Quantity doesn't buy quality syntax with neural language models. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5831–5837.
- Association for Computational Linguistics, Hong Kong, China, <https://www.aclweb.org/anthology/D19-1592>.
- Shain, C., Elsner, M., 2019. Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. 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), pp. 69–85.<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. *J. Stat. Softw.* 39 (5), 1–13.<http://www.jstatsoft.org/v39/105/>.
- Smolensky, P., Goldrick, M., 2016. Gradient symbolic representations in grammar: the case of French liaison. *Rutgers Optimality Archive* 1552, Rutgers University, pp. 1–37.
- Smolensky, P., Rosen, E., Goldrick, M., 2019. Learning a gradient grammar of French liaison. Proceedings of the 2019 Annual Meeting on Phonology, pp. 1–12.
- Team, R.C., 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.<https://www.R-project.org/>.

- van de Vijver, R., Baer-Henney, D., 2014. Developing biases. *Front. Psychol.* 5, 634.<https://www.frontiersin.org/article/10.3389/fpsyg.2014.00634>.
- 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, pp. 207–220. GLSA.
- Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. (B)* 73 (1), 3–36.