Generative Adversarial Phonology: Modeling unsupervised phonetic and phonological learning with neural networks

Gašper Beguš
University of Washington
begus@uw.edu

August 1, 2019

Abstract

This paper proposes a model of unsupervised phonetic and phonological learning of acoustic speech data based on Generative Adversarial Neural Networks. The Generative Adversarial architecture is uniquely appropriate for modeling phonetic and phonological learning because the network is trained on unannotated raw acoustic data and learning is unsupervised without any language-specific assumptions or pre-assumed levels of abstraction. The result is a Generator network that, as the paper argues, learns conditional allophonic distributions, produces innovative outputs consistent with linguistic behavior, and learns to use latent space as an approximation to phonetic and phonological features. A Generative Adversarial Network for acoustic data proposed by Donahue et al. (2019) was trained on an allophonic distribution in English, where voiceless stops surface as aspirated word-initially before stressed vowels except if followed by a sibilant [s]. The model successfully learns the allophonic alternation: the network's generated speech signal contains the conditional distribution of aspiration duration. Additionally, the network generates innovative outputs for which no evidence is available in the training data, suggesting that the network segments continuous speech signal into units that can be productively recombined. The paper also proposes a technique for establishing the network's internal representations. We identify latent variables that directly correspond to presence of [s] in the output. By manipulating these variables, we actively control the presence of [s], its frication amplitude, and spectral shape of the frication noise in the generated outputs. This suggest that the network learns to use latent variables as an approximation of phonetic and phonological features, which can thus be modeled as emergent from learning in the Generative Adversarial architecture. Crucially, we observe that the dependencies learned in training extend beyond the training range, which allows for additional exploration of learning representations. The results demonstrate that Generative Adversarial Networks bear potential for modeling phonetic and phonological learning with many further applications. The paper also discusses how the model's architecture and innovative outputs resemble and differ from linguistic behavior in language acquisition, speech disorders, and speech errors.

keywords: artificial intelligence, neural networks, generative adversarial networks, speech, phonetic learning, phonological learning, voice onset time, allophonic distribution

1 Introduction

How to model language acquisition is among the central questions in linguistics and cognitive science in general. Acoustic speech signal is the main input for hearing infants acquiring language. By the time acquisition is complete, humans are able to decode and encode information from or to a continuous speech stream and construct grammar that enables them to do so (Saffran et al.,

1996, 2007; Kuhl, 2010). In addition to syntactic, morphological, and semantic representation, the learner needs to learn phonetic representations and phonological grammar: to analyze and in turn produce speech as a continuous acoustic stream composed of discrete mental units called *phonemes*. Phonological grammar manipulates these discrete units and derives surface forms from stored lexical representations. The goal of linguistics and more specifically, phonology, is to explain how language-acquiring children construct phonological grammar, how the grammar derives surface outputs from inputs, and what aspects of the grammar are language-specific in order to tease them apart from those aspects that can be explained by general cognitive processes or historical developments (de Lacy, 2006; de Lacy and Kingston, 2013; Moreton, 2008; Moreton and Pater, 2012a,b; Beguš, 2018b).

Computational models have been invoked for the purpose of modeling language acquisition and phonological grammar ever since the rise of computational methods and computationally informed linguistics (for an overview of the literature, see Alderete and Tupper 2018a; Dupoux 2018; Jarosz 2019; Pater 2019). Modeling phonetic and phonological learning is an inherently complex task: the ideal model would need to learn articulatory representations from unannotated acoustic data on the phonetic level together with underlying representations and derivations (mappings from inputs to outputs) on the phonological level. One of the major shortcomings of the majority of the existing proposals is that learning is modeled with an already assumed level of abstraction (Dupoux, 2018). In other words, most of the proposals model phonological learning as symbol manipulation of discrete units that operates already on the abstract, discrete phonological level. The models thus require a strong assumption that phonetic learning had already taken place, and that phonemes as discrete units had already been inferred from continuous speech data (for overview of the literature, see Oudeyer 2005, 2006; Dupoux 2018). In this paper, we propose a model that combines phonetic and phonological learning: phonological distributions and features are learned simultaneously with phonetic learning of raw unannotated acoustic speech data in an unsupervised manner.

1.1 Background

As already mentioned, *phonemes* (such as /p/, /ɪ/, and /t/ in /pɪt/ 'pit') are abstract discrete mental units, the smallest meaning-distinguishing units of language (Dell et al., 1993; Kawamoto et al., 2015). A string of phoneme constitutes a morpheme, the smallest meaning-bearing unit. Phonemes are represented as feature matrices: sets of binary contrastive features (Clements, 1985; Hayes, 2009). For example, the phoneme /p/ is represented as [−sonorant, −continuant, +labial, −voice]. This feature matrix uniquely selects the phoneme /p/ from the inventory of English phonemes to the exclusion of other phonemes. *Phonological grammar* manipulates such features and feature matrices. For example, /p/ is an abstract unit that can surface (be realized) with variations on the phonetic level. English /p/ is realized as aspirated [p^h] (produced with a puff of air) word-initially before stressed vowels, but as unaspirated plain [p] (without the puff of air) if [s] immediately precedes it. This distribution is completely predictable and derivable with a simple rule (Iverson and Salmons, 1995), which is why the phoneme as an abstract mental unit is unspecified for aspiration (or absence thereof) in the underlying representation. Aspiration is represented as feature [±spread glottis]. A simple rule of the rule-based phonology (Chomsky and Halle, 1968) in (1) below can derive surface forms with or without the aspiration from underlying representation.[1]

[1]The account of aspiration in English is simplified for the purpose of this paper, because the model is trained on simplified conditions. For further details, see Iverson and Salmons (1995); Vaux and Samuels (2005).

Input	/ˈpɪt/	/ˈspɪt/	
Derivation	ˈp^{h}ɪt	∅	aspiration rule
Output	[ˈp^{h}ɪt]	[ˈspɪt]	

Table 1: Derivation of /ˈpɪt/ and /ˈspɪt/ in the rule-based approach.

$$\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ -\text{voice} \end{bmatrix} \rightarrow [+\text{spread glottis}] \,/\#\_\_\_\, [+\text{stress}] \tag{1}$$

For example, lexically stored input strings of phonemes such as /ˈpɪt/ 'pit' and /ˈspɪt/ 'spit' are unspecified for aspiration. The rule in (1) loops over the input strings and assigns [+spread glottis] value if the condition #__[+stress] is met (i.e. when a segment, represented with an underline, is immediately preceded by a word boundary # and followed by [+stress]). The surface phonetic outputs after the phonological derivation had taken place are thus [ˈp$^{\mathbf{h}}$ɪt] with the aspiration and [ˈspɪt] without the aspiration. Table 1 illustrates the derivation. The most prominent computational approach to the rule-based phonology operate with finite state automata Heinz (2010); Chandlee (2014) and constraint the classes of automata involved in phonological computation. All models within this framework operate on a symbolic level with automata that manipulate discrete units.

Phonetically, this rule is explained in Kim (1970). The spreading of the glottis onsets during [s] in sT clusters. By the time the stop is released, the glottis contracts and the aspiration ceases, which results in absence of aspiration after the release of the stop. Even if phonologically the rule in (1) is the consequence of both /s/ and /p, t, k/ being underlyingly [+spread glottis], which is why in the cluster sT the feature is not realized on the stop (Iverson and Salmons, 1995; Vaux and Samuels, 2005), the learner still needs to acquire this non-automatic allophonic distribution from speech signal.

One of the main objections of phonological theory is to explain how the grammar derives surface *outputs*, i.e. phonetic signals, from *inputs*, i.e. phonemic representations. Two influential proposals have been in the center of this discussion, the rule-based approach (outlined thus far and summarized in Table 1) and the Optimality Theory (outlined below).

The main objection against the rule-based approach to phonology is that rules are too powerful and overgenerate (Odden, 2013). In other words, rule-based phonology can derive any output from a given input by applying multiple ordered rules in the derivation (e.g. a set of simple ordered rules can turn an input /ˈpit/ into the output [ˈʒkˈæŋ] and infinite other outputs). Phonological typology, on the other hand, is considerably more limited. Moreover, modeling learning and phonological variation within the rule-based approach faces some crucial challenges (overview of the discussion in Hale and Reiss 2008; Albright and Hayes 2011; Heinz 2011).

As a response to the rule-based approach and the problem of learnability and overgeneration, a connectionist approach called Optimality Theory (Prince and Smolensky, 1993/2004) and related proposals such as Harmonic Grammar and Maximum Entropy (MaxEnt) grammar were proposed (Legendre et al., 1990; Goldwater and Johnson, 2003; Legendre et al., 2006; Wilson, 2006; Hayes and Wilson, 2008; Pater, 2009; Hayes and White, 2013; White, 2014, 2017). These models were heavily influenced by the early advances in neural network research (Alderete and Tupper, 2018a; Pater, 2019). The main advantage of Optimality Theoretic architecture is that phonological computation is modeled as optimization of outputs based on inputs. Optimality Theory introduces constraints: functions that evaluate outputs or input-output pairs. Any given input has a set of potential outputs. The winning output is chosen based on constraint violations: the output that violates

/ˈpɪt/	*#[−spread glottis][+stress] $w = 2$	IDENT-IO $w = 1.5$	H
[ˈpɪt]	−1		−2
☞ [ˈpʰɪt]		−1	−1.5

Table 2: A tableau illustrating output-input optimization in Harmonic Grammar. Each constraint assigns violations (negative integers). The Harmony score (H) is calculated from these violations and corresponding weights (w).

the lowest-weighted constraints is the winning candidate. For example, instead of deriving outputs from the input via rules, output [ˈpʰɪt] is chosen over a competing candidate [ˈpɪt] (for input /ˈpɪt/) because it satisfies the constraint stating that word-initial sequences of #[−sonorant, −continuant, −voice, −**spread glottis**] [+stress] are dispreferred. On the other hand, output [ˈspɪt] is chosen over a competing candidate [ˈspʰɪt] (for input /ˈspɪt/), because winning candidates tend to replicate inputs (the so-called faithfulness constraints, marked as IDENT-IO). The input-output optimization is formalized via Harmony scores (H). Constraints are functions that assign negative integers if an output or input-output pair incurs a constraint violation. Each constraint (C_i) has a weight (w_i). Harmony scores of output-input pairs (H(output, input)) are calculated as a sum of the product of constraint violations and their corresponding weights (Equation 2). The output candidate with the highest score is chosen as the winner (marked with ☞). Table 2 illustrates calculation of harmony scores based on constraint violations and their corresponding weights.

$$H(y, x) = \sum_{i=1}^{m} w_i C_i(y, x), \quad \text{where } y = \text{output and } x = \text{input} \tag{2}$$

In other words, phonological computation is modeled as input-output mapping based on two competing forces (formalized as constraints): the tendency to satisfy some surface form requirement and the tendency to be faithful (as identical as possible) to the input. Unlike rule-based approach, Optimality Theory is substantially more restrictive: some processes are predicted to be unattested. The second advantage of Optimality Theoretic approaches to phonology is that the model provides a theory of learnability and derives non-categorical processes (phonological variation). Harmony scores can be transformed into probability distributions ($P(y|x)$) over output candidates. In other words, every output candidate is assigned some probability of surfacing as the output, directly derivable from the Harmony score (H) in Equation 3 (Goldwater and Johnson, 2003).

$$P(y|x) = \frac{e^{H(x,y)}}{\sum_{y \in Y(x)} e^{H(x,y)}}, \quad \text{where } y = \text{output and } x = \text{input} \tag{3}$$

In the most standard version of MaxEnt and Optimality Theoretic approaches to phonology, constraints are predetermined in language acquisition (or at least constraint templates that can be filled with feature matrices are; Hayes 1999). The main task of the learner is thus to learn constraint weights. This problem is computationally most successfully addressed within the Maximum entropy model (or a multinomial logistic regression with constraints as predictors) approach.[2] The implementation, first proposed by Goldwater and Johnson (2003) has seen success in deriving

[2]For alternative proposals, such as the Gradual Learning Algorithm, see (Boersma and Hayes, 2001).

phonological learning and gradient phenomena in phonology. Learning constraint weights is thus an optimization problem that can be solved with any appropriate optimization algorithm (Pater, 2019). Several works directly compare and parallel MaxEnt grammar with experimentally observed human behavior (Wilson, 2006; White, 2014, 2017; Moreton et al., 2017). Another advantage of this model is that learning biases and asymmetries in rate of learning can be encoded in the computational model. Constraints can have non-zero prior weights and learning rate can be encoded as prior variance Wilson (2006) or prior means White (2017) in regularization term.

1.2 Neural networks

The weighted-constraint approaches to phonology including the MaxEnt grammar approach (as a multinomial linear logistic regression model) are in many ways related to neural networks (Smolensky and Legendre, 2006; Alderete and Tupper, 2018a; Pater, 2019). Modeling linguistic data with neural networks has seen a rapid increase in the past few years (Alderete et al. 2013; Avcu et al. 2017; Alderete and Tupper 2018a; Mahalunkar and Kelleher 2018; Weber et al. 2018; Dupoux 2018; Prickett et al. 2019, for cautionary notes, see Rawski and Heinz 2019). While the MaxEnt grammar as well as the rule-based approaches require language-specific devices (such as constraints or rules, binary features, discrete mental units of representation etc.), one of the promising implications of the neural network modeling is the ability to test generalizations that the models produce without language-specific assumptions (Pater, 2019).

The majority of existing computational models in phonology (both using the MaxEnt and neural network methods), however, model learning as symbol manipulation and operate with discrete units—either with completely abstract made-up units or with discrete units that feature some phonetic properties and can be approximated as phonemes. This means that either the phonetic and phonological learning are modeled separately or one is assumed to have already been completed (Martin et al., 2013; Dupoux, 2018). This is true for both proposals that model phonological distributions or derivations (Alderete et al., 2013; Prickett et al., 2019) and featural organizations (Faruqui et al., 2016; Silfverberg et al., 2018). Relatively fewer proposals that model continuous phonetic data also assume at least some level of abstraction and operate with already extracted features (e.g. formant values) on limited "toy" data (e.g. Pierrehumbert 2001; Kirby and Sonderegger 2015 for a discussion, see Dupoux 2018). Guenther and Vladusich (2012), Guenther (2016) and Oudeyer (2001, 2002, 2005, 2006) propose models that use simple neural maps that are based on actual correlates of neurons involved in speech production in the human brain (based on various brain imaging techniques). Their models, however, do not operate with raw acoustic data (or require extraction of features in a highly abstract model of articulators; Oudeyer 2005, 2006), require a level of abstraction in the input to the model, and do not model phonological processes — conditional allophonic distributions. Phonological learning in most of these proposals is thus modeled as if phonetic learning (or at least a subset of phonetic learning) had already taken place: the initial state already includes phonemic inventories, phonemes as discrete units, feature matrices that had already been learned, or extracted phonetic values.

Prominent among the few models that operate with raw phonetic data are Gaussian mixture models for category learning or phoneme extraction (Schatz et al., 2019; Lee and Glass, 2012). Schatz et al. (2019) propose a Dirichlet process Gaussian mixture model that learns categories from raw acoustic input in an unsupervised learning task. The model is trained on English and Japanese data and the authors show that the asymmetry in perceptual [l]∼[r] distinction between English and Japanese falls out automatically from their model. The primary purpose of the model in Schatz et al. (2019) is modeling perception and categorization: they model how a learner is able to categorize raw acoustic data into sets of discrete categorical units that have phonetic values

(i.e. phonemes). No phonological processes are modeled in the proposal.

Recently, neural network models for unsupervised feature extraction have seen success in modeling acquisitions of phonetic features from raw acoustic data. The model in Shain and Elsner (2019), for example, is an autoencoder neural network that is trained on pre-segmented acoustic data. The model takes as an input segmented acoustic data and outputs values that can be correlated to phonological features. Learning is, however, not completely unsupervised as the network is trained on pre-segmented phones. Thiollière et al. (2015) similarly propose an architecture that extracts units from unsupervised speech data. Other proposals for unsupervised acoustic analysis with neural network architecture are similarly primarily concerned with unsupervised feature extraction (Kamper et al., 2015).

These proposals, however, do not model learning of phonological distributions, but only of feature representations, and crucially are not generative, meaning that the models do not output innovative data, but try to replicate the input as closely as possible (e.g. in the autoencoder architecture). As will be argued below, the model based on Generative Adversarial network learns not only to generate innovative data that closely resemble human speech, but also learns internal representations that directly resemble phonological features simultaneously with unsupervised phonetic learning from raw acoustic data. Additionally, the model is generative and outputs both the conditional allophonic distributions in the data and innovative data that can be compared to productive outputs in human speech acquisition. To the author's knowledge, this is the first proposal that uses GAN architecture to model generative phonetic and phonological learning.

1.3 A Generative Adversarial model of phonology

The advantage of the GAN architecture is that learning is completely unsupervised and that, as is argued in Section 3 below, phonetic learning is simultaneous with phonological learning. The discussion on the relationship between phonetics and phonology is highly complex (Kingston and Diehl, 1994; Cohn, 2006; Keyser and Stevens, 2006). Several opposing proposals, however, argue that the two interact at various different stages and are not dissociated from each other (Hayes, 1999; Pierrehumbert, 2001; Fruehwald, 2016, 2017). A network that models learning of phonetics from raw data and shows signs of learning discrete phonological units at the same time is likely one step closer to reality than models that operate with symbolic computation and assume phonetic learning had already taken place and is independent of phonology and vice versa. Additionally, the GAN architecture models the production-perception loop in phonetics and phonology that other models generally lack. The Generator's outputs can be interpreted as the basis for articulatory targets in human speech that are sent to articulators for execution. The latent variables in the input of the Generator can be modeled as articulatory parameters that the Generator learns to output into a speech signal by attempting to maximize the error rate of a Discriminator network that distinguishes between real data and generated outputs. The Discriminator network has a clear parallel in human speech perception, production, and acquisition: the imitation principle (Nguyen and Delvaux, 2015). The Discriminator's function is to enforce the Generator's outputs to be as similar to the input as possible. The GAN network thus incorporates both the pre-articulatory production elements (the Generator) as well as the perceptual element (the Discriminator) in speech acquisition.

We train a Generative Adversarial Network architecture implemented for audio files in Donahue et al. (2019) (WaveGAN) on continuous raw speech data that contains information for an allophonic distribution: word-initial pre-vocalic aspiration of voiceless stops ([ˈpʰɪt] ∼ [ˈspɪt]). The data is curated in order to control non-desired effects, which is why only sequences of the shape #TV and #sTV are fed to the model. This allophonic distribution is uniquely appropriate for testing

learnability in a GAN setting, because the dependency between the presence of [s] and duration of VOT is not strictly local. To be sure, the dependency is local in phonological terms, as [s] and T are two segments and immediate neighbors, but in phonetic terms, a period of closure intervenes between the aspiration and the period (or absence thereof) of frication noise of [s]. It is not immediately clear whether a GAN model is capable of learning such non-local dependencies.

The hypothesis of the computational experiment presented in Section 3 is the following: if VOT duration is conditioned on the presence of [s] (i.e. there is significant difference between the two groups) in output data generated from noise by the Generator network, it means that the Generator network has successfully learned a phonetically non-local allophonic distribution. This distribution is not automatic in English, which means that not only phonetic, but also phonological distributions are modeled with this approach.

The results suggest that phonetic and phonological learning can be modeled simultaneously and in unsupervised mode directly from what language acquiring infants are exposed to: raw acoustic data. A GAN model trained on an allophonic distribution is successful in learning to generate acoustic output from random noise. The generated acoustic outputs include evidence that the Generator network learns the conditioned distribution of VOT duration. Additionally, the model outputs innovative data for which no evidence was available in the training data, allowing a direct comparison between human speech data and GAN's generated output. As argued in Section 3.3, some outputs are consistent with human linguistic behavior and suggest that the model learns phones as discrete units that can be recombined, directly resembling phonemic representation and productivity in human language acquisition (Section 4).

The paper also proposes a technique for establishing the Generator's internal representations. What neural networks actually learn is a challenging question with no easy solutions. The inability to uncover network's representations has been used as an argument against neural network approaches to linguistic data (Rawski and Heinz, 2019). We argue that internal representation of a network can be, at least partially, uncovered. By fitting annotated outputs of the Generator network and the latent space of the network to logistic regression models, we identify values in the latent space that correspond to linguistically meaningful features in generated output. The paper demonstrates that manipulating the chosen values in the latent space have clear phonetic and phonological effects in the generated outputs, such as presence of [s] and the amplitude of its frication. In other words, the GAN network learns to use random noise as phonetic and phonological features. The paper proposes that dependencies, learned during training in a latent space that is limited by some interval extend beyond that interval. This crucial step allows for discovery of several phonetic properties that the model learns. We argue that phonological features thus emerge automatically from the GAN architecture.

By modeling phonetic and phonological learning with neural networks without any language specific assumptions, the paper also addresses a broader question of how much language-specific elements we need in models of grammar and language acquisition. Most of the existing models require at least some language-specific devices, such as rules in rule-based approach or pre-determined constraints with features and feature matrices in connectionist approaches. The model proposed here lacks any language-specific device. Comparing performance of such model with competing proposals and human behavior should result in a better understanding of what aspects of phonological grammar and acquisition are domain-specific (Section 4).

2 Materials

2.1 The model

Generative Adversarial Networks, proposed by Goodfellow et al. (2014), have seen a rapid expansion in a variety of tasks, including but not limited to computer vision and image generation (Radford et al., 2015). The main characteristic of GANs is the architecture that involves two networks: the Generator network and the Discriminator network (Goodfellow et al., 2014). The Generator network is trained to generate data from random noise, while the Discriminator is trained on distinguishing real data from the outputs of the Generator network (Figure 1). The Generator is trained to generate data that minimizes accuracy of the Discriminator network. The training results in a Generator (G) network that takes random noise as its input (e.g. multiple variables with uniform distributions) and outputs data such that the Discriminator is inaccurate in distinguishing the generated from real data. Goodfellow et al. (2014) summarizes the architecture (repeated here in Equation 14), where V is value function that the Generator maximizes and Discriminator minimizes, G is Generator, D is Discriminator, x is data from $P_{\text{data}}(x)$, z are latent input variables from prior P_z.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log(D(x)] + \mathbb{E}_{z\sim P_z(z)}[1-\log(D(G(z)))] \tag{4}$$

Applying the GAN architecture on a time-series data such as continuous speech stream faces several challenges. Recently, Donahue et al. (2019) proposed an implementation of a Deep Convolutional Generative Adversarial Network proposed by Radford et al. (2015) for audio data (WaveGAN). The model takes one-second long raw audio files as inputs, sampled at 16 kHz with 16-bit quantization. The audio files are converted into a vector and fed to the Discriminator network as real data. Instead of the two-dimensional 5×5 filters, WaveGAN model uses one-dimensional 1×25 filters and larger upsampling. The main architecture is preserved as in DCGAN, except that an additional layer is introduced in order to generate longer samples. The Generator network takes as input z, a vector of one hundred uniformly distributed variables ($z\sim\mathcal{U}(-1,1)$) and outputs 16,384 data points, which constitutes the output audio signal. The network has five 1D convolutional layers (Donahue et al., 2019). The Discriminator network takes 16,384 data points (raw audio file) as its input and outputs a single logit. The initial GAN design as proposed by Goodfellow et al. (2014) trained the Discriminator network on distinguishing real from generated data. Training such models, however, faced substantial challenges (Donahue et al., 2019). Donahue et al. (2019) implements WGAN-GP strategy (Arjovsky et al., 2017; Gulrajani et al., 2017), which means that the Discriminator is trained "as a function that assists in computing the Wasserstein distance" (Donahue et al., 2019). The WaveGAN model (Donahue et al., 2019) uses ReLU activation in all but the last layer for the Generator network, and Leaky ReLU in all layers in the Discriminator network (as recommended for DCGAN in Radford et al. 2015). The model is implemented in TensorFlow 1.13 (Abadi et al., 2015) in Donahue et al. (2019). For exact dimensions of each layer and other details of the model, see Donahue et al. (2019).

2.2 Training data

The model was trained on allophonic distribution of voiceless stops in English. As already mentioned in Section 1, voiceless stops /p, t, k/ surface as aspirated (produced with a puff of air) [p^h, t^h, k^h] in English in word-initial position when immediately followed by a stressed vowel (Lisker, 1984; Iverson and Salmons, 1995; Vaux, 2002; Vaux and Samuels, 2005; Davis and Cho, 2006). If an alveolar sibilant [s] precedes the stop, however, the aspiration is blocked and the stop surfaces

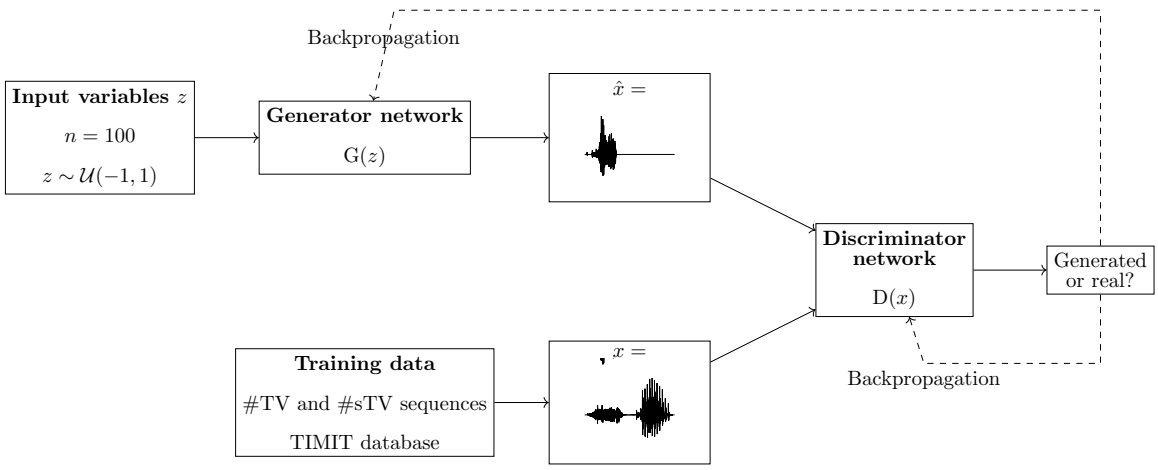

Figure 1: A diagram showing the Generative Adversarial architecture as proposed in Goodfellow et al. (2014); Donahue et al. (2019) and trained on data from the TIMIT database in this paper.

as unaspirated [p, t, k] (Lisker, 1984). A minimal pair illustrating this allophonic distribution is ['pʰɪt] 'pit' vs. ['spɪt] 'spit'. The most prominent phonetic correlate of this allophonic distribution is the difference in Voice Onset Time (VOT) duration (Abramson and Whalen, 2017) between the aspirated and unaspirated voiceless stops.

The model was trained on data from the TIMIT database (S Garofolo et al., 1993).[3] The corpus was chosen because it is one of the largest currently available hand-annotated speech corpora, the recording quality is relatively high, and the corpus features a relative high degree of variability. The database includes 6300 sentences, 10 sentences per 630 speakers from 8 major dialectal areas in the US (S Garofolo et al., 1993). The training data consist of 16-bit .wav files with 16 kHz sampling rate of word initial sequences of voiceless stops /p, t, k/ (= T) that were followed by a vowel (#TV) and word initial sequences of /s/ + /p, t, k/, followed by a vowel (#sTV). The training data includes 4,930 sequences with the structure #TV and 533 sequences with the structure #sTV (5,463 total). Figure 2 illustrates typical training data: raw audio files with speech data, but limited to two types of sequences, #TV and #sTV. Figure 2 also illustrates that the duration of VOT depends on a condition that is not immediately adjacent in phonetic terms: absence/presence of [s] is interrupted from the VOT duration by a period of closure in the training data.

Both stressed and unstressed vowels are included in the training data. Including both stressed and unstressed vowels is desirable, as this condition crucially complicates learning and makes the task for the neural network more challenging. Aspiration is less prominent in word-initial stops not followed by a stressed vowel. This means that in the condition #TV, the stop will be either fully aspirated (if followed by a stressed vowel) or unaspirated (if followed by a unstressed vowel). In the #sTV condition, the stop is never aspirated. Learning of two conditions is more complex if the dependent variable in one condition can range across the variable in the other condition.

To confirm the presence of this durational distribution in the training data, VOT durations based on TIMIT's hand annotations were measured across the two conditions. VOT is in TIMIT annotated from the release of the stop to the onset of the following vowel. Slices for which no VOT duration exists (only closure duration that includes the VOT) were excluded from this analysis,

[3]Donahue et al. (2019) train the model on SC09 and TIMIT databases, but the results are not useful for modeling phonological learning, because the model is trained on continuous speech stream and the generated sample fail to produce analyzable results for phonological purposes.

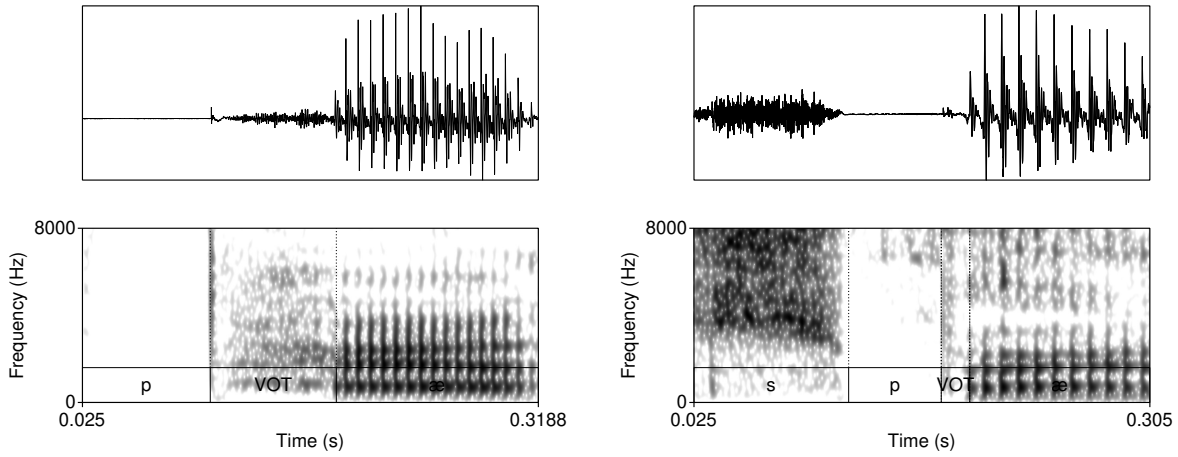

Figure 2: Waveforms and spectrograms ($0 - 8000$ Hz) of [p^{h}æ] (left) and [spæ] (right) illustrating typical training data with annotations from TIMIT. Only the raw audio data (in .wav format) were used in training. The annotation illustrates a substantially longer duration of VOT in word-initial stops when no [s] precedes.

Structure	**Place**	**VOT**	**SD**	**Lowest**	**Highest**
	p	49.6	18.0	7.3	115.5
#TV	t	55.2	20.7	9.8	130.0
	k	67.5	19.5	12.5	153.1
	p	19.4	7.1	9.4	49.2
#sTV	t	25.6	7.9	10.6	65.0
	k	30.1	8.6	14.4	55.0

Table 3: Raw VOT durations in ms for the training data with SD and Range.

but were included in the training: altogether 47 sequences were thus excluded. While the TIMIT database is occasionally misaligned, the errors are minor and likely do not crucially affect the outcomes. Table 3 and Figure 3 summarize raw VOT durations across three places of articulation. Speaker identity is not included in the model, because it is irrelevant for the purpose of training a GAN network.

To test significance of the presence of [s] as a predictor of VOT duration, the data were fit to a linear model with two predictors: Structure (presence vs. absence of [s]) and Place of articulation of the target stop (with three levels — [p], [t], [k]) and their interaction. Structure was treatment-coded (with absence of [s] as the reference level), while place of articulation of the stop was sum-coded (with [k] as reference). The interaction term is significant ($F(2) = 6.97, p < 0.001$), which is why it is kept in the final model. The model shows that at the mean of the Place of articulation as a predictor, VOT is approximately 32.4 ms shorter if T is preceded by [s]. The 95% confidence intervals for this difference are [−34.3 ms, −30.6 ms]. Figure 4 illustrates the significant difference and its magnitude between the two conditions across the three places of articulation. The significant interaction #sTV:[t] is not informative and irrelevant for our purposes.

The training data is not completely naturalistic: only #TV and #sTV sequences are sliced from continuous speech data. This, however, has a desirable effect. The primary purpose of this paper is to test whether a GAN model can learn an allophonic distribution from data that consists of raw

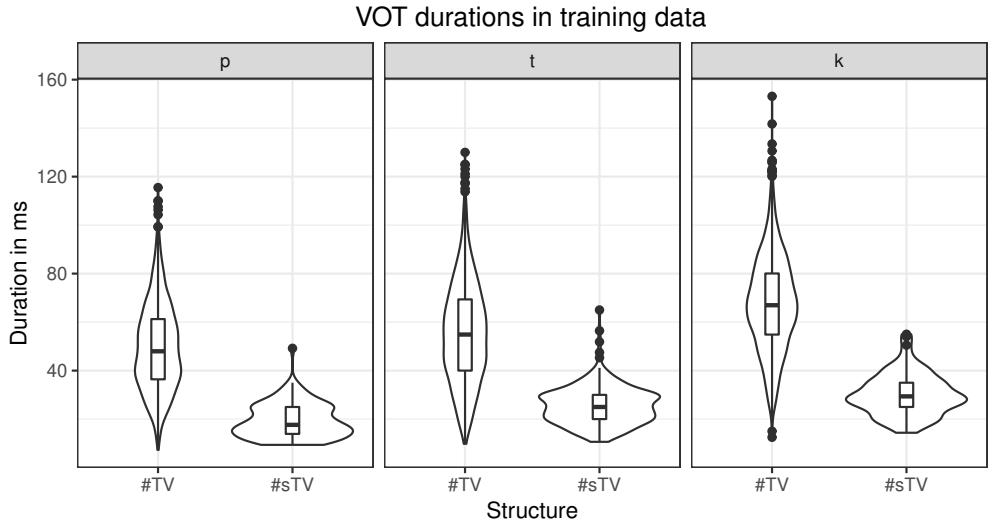

Figure 3: Violin plots with box-plots of durations in ms of VOT in the training data based on two conditions: when word-initial TV sequence is not preceded by [s] (#sTV) and when it is preceded by [s] (#sTV) accross the three places of articulation: [p], [t], [k].

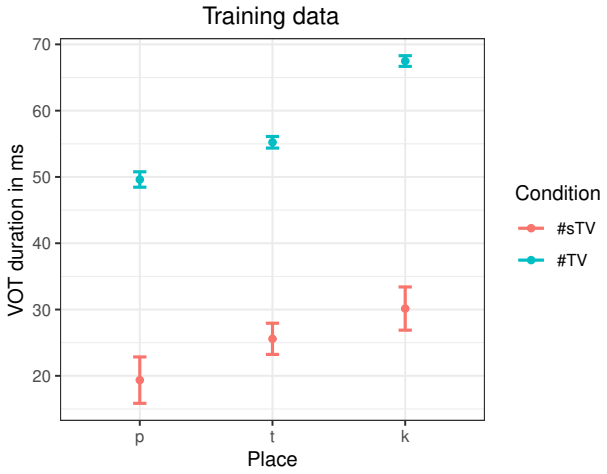

Figure 4: Distribution of VOT durations as estimated from a linear model.

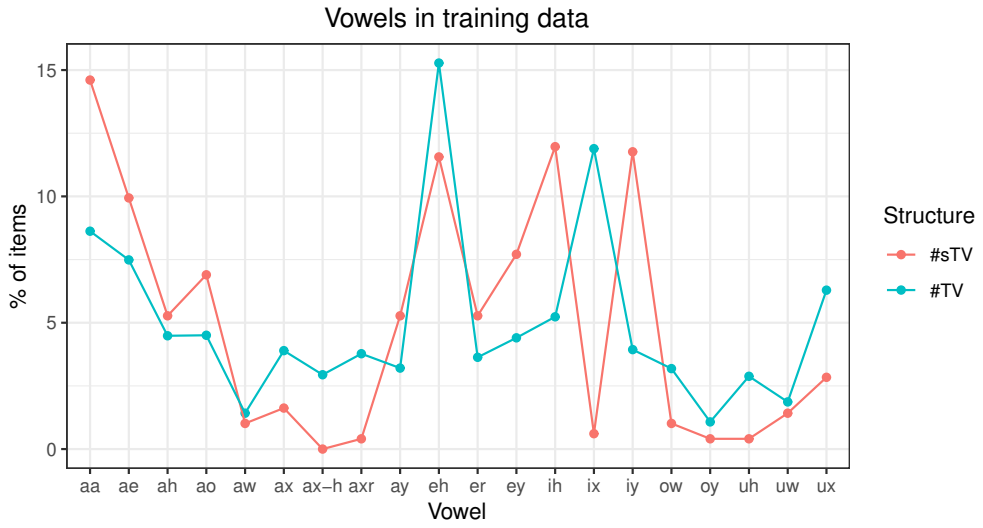

Figure 5: Distribution of training items according to vowel identity as described in TIMIT in ARPABET, where aa = ɑ, ae = æ, ah = ʌ, ao = ɔ, aw = aʊ, ax = ə, ax-h = ə̥, axr = ɚ, ay = aɪ, eh = ɛ, er = ɝ, ey = eɪ, ih = ɪ, ix = ɨ, iy = i, ow = oʊ, oy = oɪ, uh = ʊ, uw = u, ux = ʉ in International Phonetic Alphabet.

acoustic inputs. If the whole lexicon were included in the training data, the distribution of VOT duration could be conditioned on some other distribution, not the one this paper is predominately interested in: presence or absence of [s]. It is thus less likely that the distribution of VOT duration across the main condition of interest, presence of [s], is conditioned on some other unwanted factor in the model precisely because of the balanced design of the training data. The only condition that can potentially influence learning is the distribution of vowels across the two conditions. Figure 5, however, shows that vowels are relatively equally distributed across the two conditions, which means that vowel identity likely does not influence the outcomes substantially. Finally, vowel duration (or the equivalent of speech rate in real data) and identity are not controlled for in the present experiment. To control for vowel duration, VOT duration would have to be modeled as a proportion of the following vowel duration. Several confounds that are not easy to address would be introduced, the main of which is that vowel identification is not unproblematic for generated inputs with fewer training steps 3.2. Because the primary interest of the experiment is the difference in VOT durations between two groups (presence and absence of [s]) and substantial differences in vowel durations (or speech rate) between the two groups are not expected, we do not anticipate the results to be substantially influenced by speech rate.

3 Experiment

3.1 Training and generation

The model was trained on a single NVIDIA K80 GPU. The network was trained at an approximate pace of 40 steps per 300 s. The purpose of this paper is to model phonetic and phonological learning. For this reason, the Generator network was not fully trained until convergence: the data was generated and examined at different points as the Generator network was in the process of being trained. Outputs of the network are analyzed at two stages: after 1,474 steps (Section 3.2)

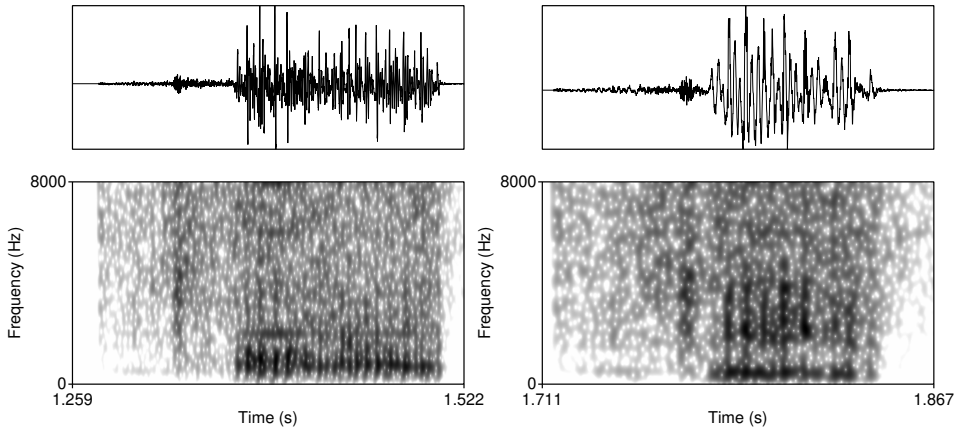

Figure 6: Waveforms and spectrograms (0–8,000 Hz) of a typical generated samples of #TV sequences from a Generator trained after 1474 steps.

when the outputs still substantially violate the input data and after 12,255 steps (Section 3.3), when the outputs of the model closely resembles human speech.

3.2 Model 1: 1,474 steps

In the first test of the model, the network was trained with 1,474 steps (approximately 86 epochs). 950 outputs of the Generator network were generated for analysis. Every generated output was listened to and spectral properties were manually observed by the author. At this point, the model performs poorly, which is why only qualitative analysis of the generated samples is possible. Nevertheless, some significant observation emerge even in this initial model.

Most of the generated samples already have a clear vocalic element with more or less pronounced formant structure and a non-vocalic element — VOT after the release of closure. Figures 6 illustrates two typical outputs with the structure #TV. The spectrograms show both vocalic structure and frication noise from aspiration. VOT duration is substantial. The Generator also generates sequences with the structure #sTV, illustrated in Figure 7. The peculiarity about the #sTV sequences at this point is that the sibilant part seems substantially shorter (with a narrow band of [s]-like frequency distribution) and the closure features relatively high amount of noise. This limited sample already suggest that the Generator might be learning the conditional VOT distribution as outputs with [s] feature no obvious VOT duration (although bursts are not clearly visible either).

At this point, the Generator network also generate samples that substantially violate distributions in the training data. One such output includes three consecutive sibilants [sss]; another includes two or three consecutive vocalic elements divided by periods of reduced noise (Figure 8). Occasionally, the order of segments is violated. The left spectrogram in Figure 7 shows that a short vocalic element surfaces between [s] and the closure. The left spectrogram in Figure 8 shows that a period of silence (marked with an arrow) intervenes during the vowel V.

3.3 Model 2: 12,255 steps

The Generator network after 12,225 steps ($\sim$ 716 epochs) generates acoustic signal that appears substantially closer to actual speech data compared to Model 1. Figure 9 illustrates a typical

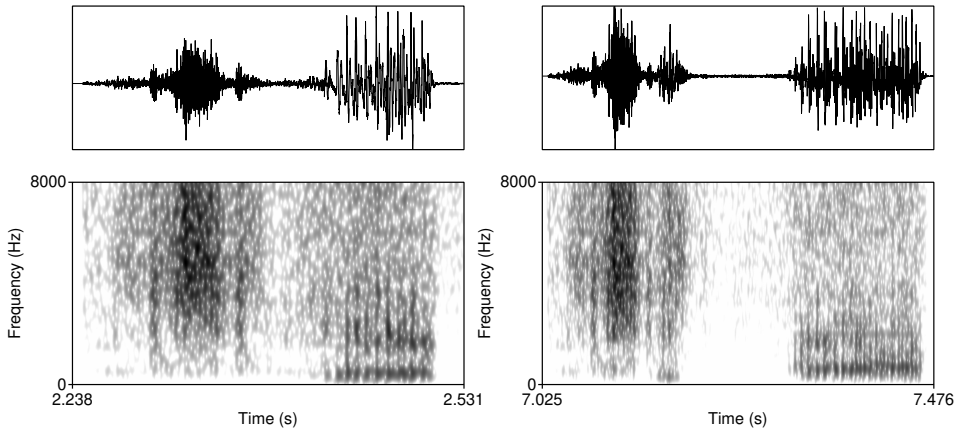

Figure 7: Waveforms and spectrograms (0–8,000 Hz) of typical generated samples of #sTV sequences from a Generator trained after 1474 steps.

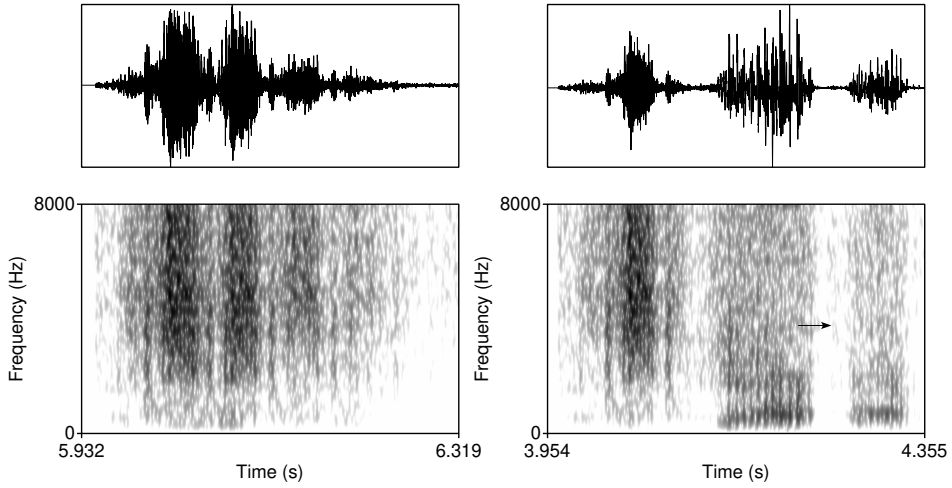

Figure 8: Waveforms and spectrograms (0–8,000 Hz) of generated samples that violate training data distributions from a Generator trained after 1474 steps. The left spectrogram shows a sequence of three [s] divided by periods of reduced frication noise. The right spectrogram illustrates silence (marked with an arrow) during the vocalic element.

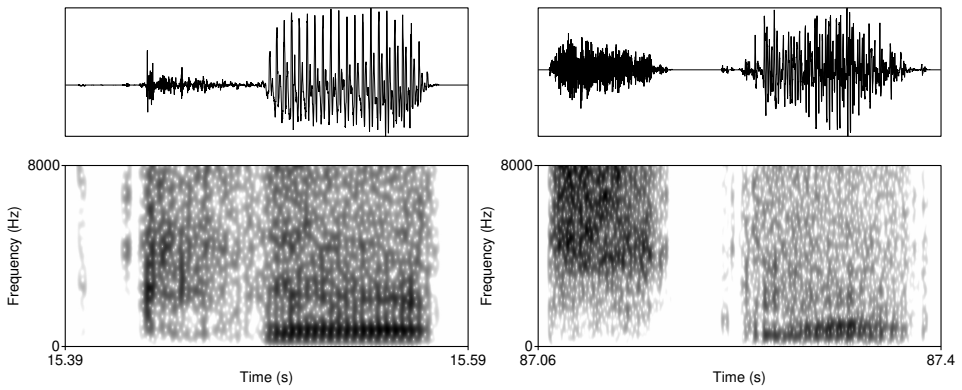

Figure 9: Waveforms and spectrograms (0–8,000 Hz) of a typical generated samples of #TV (left) and #sTV (right) sequences from a Generator trained after 12,255 steps.

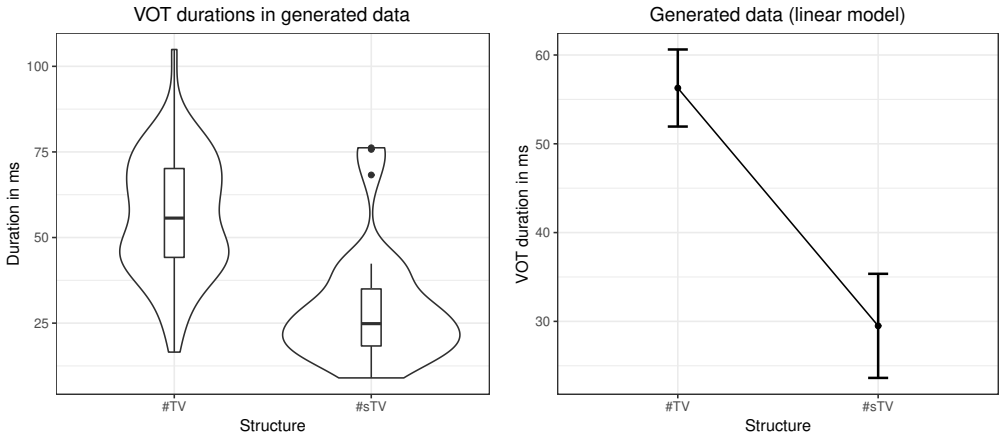

Figure 10: (left) Violin plots with box-plots of durations in ms of VOT in the generated data based on two conditions: when word-initial TV sequence is not preceded by [s] (#sTV) and when it is preceded by [s] (#sTV). (right) Estimates of VOT duration with 95% confidence intervals across two conditions, #TV and #sTV in the generated data for a model trained after 12,255 steps.

generated sample of #TV (left) and #sTV (right) structures with a substantial difference in VOT durations.

To test whether the Generator learns the conditional distribution of VOT duration, the Generated samples were annotated for VOT duration. VOT duration was measured from the release of closure to the onset of periodic vibration with clear formant structure. Altogether 96 generated samples were annotated, 62 in which no period of frication of [s] preceded and 34 in which [s] precedes the TV sequence. Only samples with structure that resembles real acoustic outputs were annotated. Figure 10 shows raw distribution of VOT durations in the generated samples that closely resembles the distribution in the training data (Figure 3).

To test significance of the observed distribution, the generated data were fit to a linear model with only one predictor: absence of [s] (Structure). Place of articulation or following vowel were not added in the model, because it is often difficult to recover place of articulation or vowel quality of generated samples. Structure is a significant predictor of VOT duration: $F(1) = 53.1, p <$

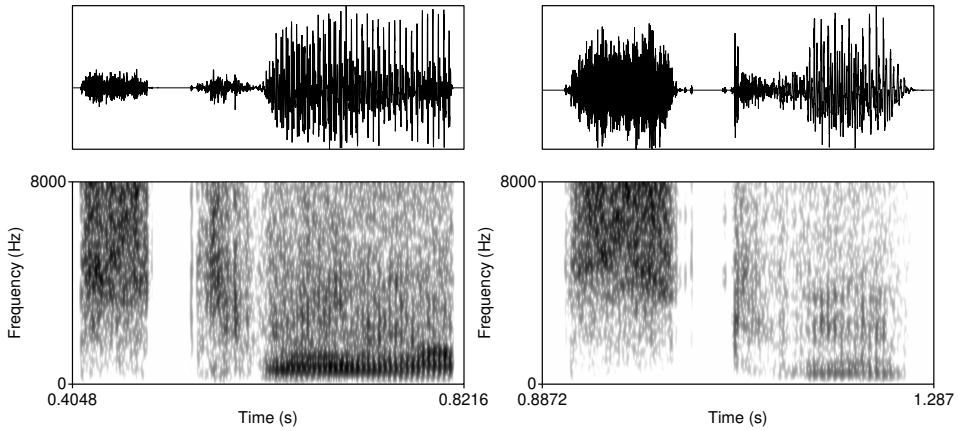

Figure 11: Waveforms and spectrograms (0–8000 Hz) of two generated outputs of #sTV sequences in which the stop has longer VOT than any VOT in #sTV condition in the training data.

0.0001. The estimates for Intercept (duration of VOT when no [s] precedes) are $\beta = 56.2$ ms, $t = 25.74, p < 0.0001$. VOT is on average 26.8 ms shorter if [s] precedes the TV sequence and this difference is significant ($\beta = -26.8$ ms, $t = -7.29, p < 0.0001$). Figure 10 illustrates estimates of VOT duration across the two conditions with 95% confidence intervals.

While VOT duration is significantly shorter if [s] precedes the #TV sequence in the generated data, the model shows clear traces that the learning is not complete and that the generator network fails to learn the distribution *categorically* at 12,255 steps. The three longest VOT durations in the #sTV condition in the generated data are 68.3 s, 75.7 s, and 76.2 s. In all three cases is the VOT longer than the longest VOT duration of any #sTV sequence in the training data (longest is 65 ms; see Table 3 and Figure 3). Figure 11 shows two such cases. This generalization holds even in proportional terms (i.e. while controlling for "speech rate"): the generated data contains the highest ratio between the VOT duration and the frication duration of [s]. The ratio VOT/[s] is 1.70 in one of the generated outputs with the VOT duration that exceeds VOT durations in the training data. In another, the ratio is 1.07. The highest ratio in the training data, on the other hand, is 0.77[4] (in an acoustically very different token compared to the generated outputs). The ratio in all other tokens in the training data are even lower, below 0.69. It is clear that the generator fails to reproduce the conditioned durational distribution from the training data in these particular cases.

Longer VOT duration in the #sTV condition in the generated data compared to training data is not the only violation of the training data that the Generator outputs and that resembles linguistic behavior in humans. Occasionally, the Generator outputs a linguistically valid #sV sequence for which no evidence was available in the training data. The minimal duration of closure in #sTV sequences in the training data is 9.2 ms, the minimal duration of VOT is 9.4 ms. All sequences containing a [s] from the training data were manually inspected by the author and none of them contain a #sV sequence without a period of closure and VOT. Homorganic sequences of [s] followed by an alveolar stop [t] (#stV) are occasionally acoustically similar to the sequence without the stop (#sV) because frication noise from [s] carries onto the homorganic alveolar closure which can be

[4]The TIMIT annotations would yield a ratio of 1.17, but the token was annotated by the author and the ratio appears much smaller. In any case, even with TIMIT's annotation, the ratio with value of 1.70 in the generated data is still substantially higher than the 1.17.

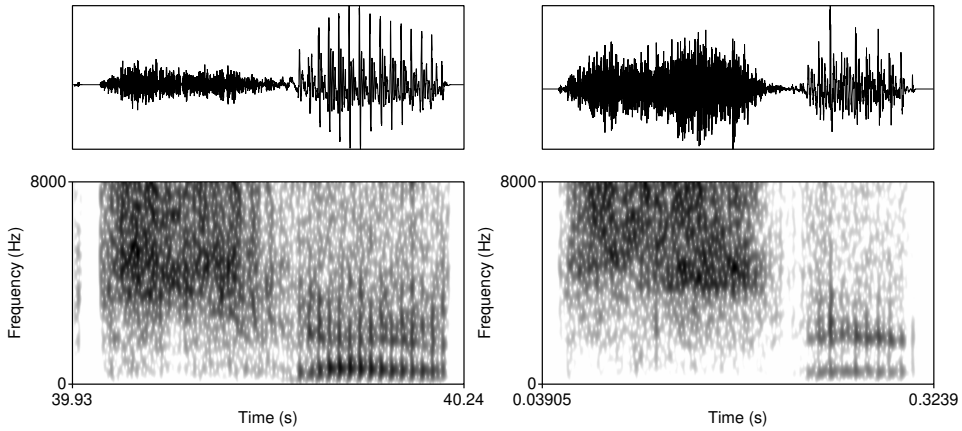

Figure 12: Waveforms and spectrograms (0–8000 Hz) of two generated outputs of the shape #sV sequences for which no evidence was present in the training data. The sample on the left was generated after 16,715 steps.

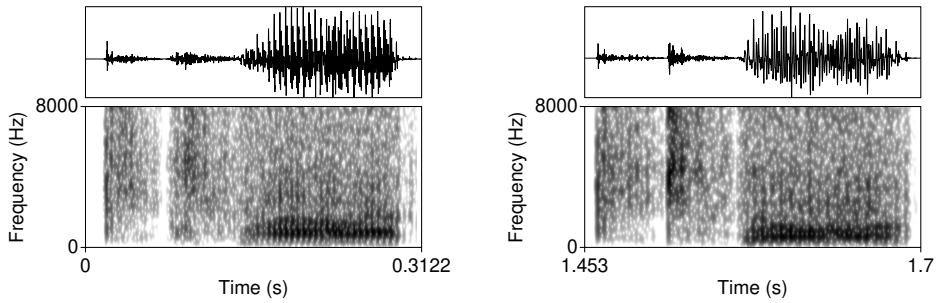

Figure 13: Waveforms and spectrograms (0–8000 Hz) of two generated outputs of the shape #TTV sequences for which no evidence was present in the training data.

very short. However, there is a clear fall and a second rise of noise amplitude after the release of the stop in #stV sequences. Figure 12 shows two cases of the Generator network outputting a #sV sequence without any stop-like fall of the amplitude. In other words, the Generator network outputs a linguistically valid sequence #sV without any evidence for existence of this sequence in the training data.

Similarly, the Generator occasionally outputs a sequence with two stops and a vowel (#TTV). To the author's knowledge, no evidence for such sequences is available in the training data. Figure 13 illustrates two such examples in which the vocalic period is preceded by two bursts, two periods of aspiration and a short period of silence between the aspiration noise of the first consonant and the burst of the second consonant that corresponds to closure of the second stop. Spectrograms show the distribution of energy differs across the two bursts and aspiration noises, suggesting that the output represents a heterogranic cluster [pt] followed by a vowel.

Measuring overfitting is a substantial problem for Generative Adversarial Networks with no consensus on the most appropriate quantitative approach to the problem (Goodfellow et al., 2014; Radford et al., 2015). The danger with overfitting in a GAN is that the Generator network would learn to fully replicate the input. Donahue et al. (2019) test overfitting on models trained with sub-

stantially higher number of steps (200,000) compared to our model (12,255) and presents evidence that GAN models trained on audio data do not overfit even with substantially higher number of training steps. The best evidence against overfitting is precisely the fact that the Generator network outputs samples that substantially violate output distributions (Figures 11 and 12).

3.4 Establishing internal representations

Establishing internal representations of a neural network is a challenging task (Lillicrap and Kording, 2019). Since the paper's primary goal is not clustering of phones (as in some of the current proposals on establishing internal representations), the methods such as Principal Component Analysis, Multidimensional Scaling (Bullinaria, 1997) are not as appropriate. We propose a different method based on logistic regression. First, 3,800 outputs from the Generator network trained after 12,255 steps were generated and manually annotated for presence or absence of [s]. 271 outputs (7.13%) were annotated as involving a segment [s]. Frication that resembled [s]-like aspiration noise after the alveolar stop and before high vowels was not annotated as involving [s].[5] Innovative outputs such as an #[s] without the following vowel or #sV sequences were annotated as involving an [s].

The annotated data together with values of latent variables for each generated sample (z) were fit to a logistic regression generalized additive model (using the *mgcv* package; Wood 2011 in R Core Team 2018) with the presence or absence of [s] as the dependent variable (binomial distribution of successes and failures) and smooth terms of latent variables (z) as predictors of interest (estimated as penalized thin plate regression splines; Wood 2011). Generalized additive model were chosen in order to avoid assumptions of linearity: it is possible that latent variables are not linearly correlated with features of interest in the output of the Generator network. The initial full model (FULL) includes smooths for all 100 variables in the latent space that are uniformly distributed with the range of $(-1, 1)$ as predictors.

The models explored here do not serve for hypothesis testing, but for exploratory purposes: to identify variables, the effects of which will be tested with a different method (see Figure 17). For this reason, several techniques to reduce the number of predictors are explored and compared: the latent variables for further analysis are then chosen based on combined results of different exploratory models.

First, we refit the model with modified smoothing penalty (MODIFIED), which allows shrinkage of the whole term (Wood, 2011). Second, we refit the model with original smoothing penalty (SELECT), but with an additional penalty for each term if all smoothing parameters tend to infinity (Wood, 2011). Finally, we identify non-significant terms by Wald test for each term (using *anova.gam()* with $\alpha = 0.05$) and manually remove them from the model (EXCLUDED). 38 predictors are thus removed.

The estimated smooths appear mostly linear. We also fit the data to a linear logistic regression model (LINEAR) with all 100 predictors. To reduce the number of predictors, another model is fit (LINEAR EXCLUDED) with those predictors removed that do not improve fit (based on the AIC criterion when each predictor is removed from the full model). 23 predictors are thus removed. The advantage of the linear model is that predictors are parametrically estimated.[6]

While the number of predictors in the models is high even after shrinkage or exclusion, there is

[5] It is possible that some outputs were mislabeled, but the probability is low and the magnitude of mislabeled data would be minimal enough not to influence the results. The author manually inspected spectrograms of all generated data.

[6] It would be possible to estimate smooth terms for only a subset of predictors, but such a model is unlikely to yield different results.

	df	AIC
FULL	108.94	1018.38
MODIFIED	88.06	1031.03
EXCLUDED	71.51	1008.20
LINEAR	101.00	1036.04
LINEAR EXCLUDED	78.00	1007.06

Table 4: AIC values of five fitted models with corresponding degrees of freedom (df), fitted with Maximum Likelihood. AIC of SELECT is not listed because it was not fitted with ML; AIC of SELECT fitted with REML is, however, similar to EXCLUDED (=1,008.46 vs. 1008.54).

little multicollinearity in the data as the 100 variables are randomly sampled for each generation. The highest Variance Inflation Factor in the linear logistic regression models (LINEAR and LINEAR EXCLUDED) estimated with *glm()* function is 1.287. All concurvity estimates in the non-linear model are below 0.3 (using *concurvity()* in Wood 2011). While the number of successes per predictor is relatively low, it is unlikely that more data would yield substantially different results (as will be shown below, the model successfully identifies those values that have direct phonetic correlates in the generated data).

Six models are thus fit in an exploratory method to identify variables in the latent space that predict presence of [s] in generated outputs. Table 4 lists AIC for each model. The LINEAREXCLUDED model has the lowest AIC score. All six models, however, yield similar results.

To identify latent variables with highest correlation with [s] in the output, we extract χ^2 estimates for each term from the generalized additive models and estimates of slopes (β) from the linear model. Figure 14 plots those values. The plot points to a substantial difference between the highest seven predictors and the rest of the latent space. Seven latent variables are thus identified (z_5, z_{11}, z_{49}, z_{29}, z_{74}, z_{26}, z_{14}) as potentially having the largest effect on presence or absence of [s] in output. Figure 15 plots smooths of the seven predictors from a non-linear model SELECT.

The same seven variables are also identified as having the highest estimates in a Lasso regression for binomial data, estimated with the *glmnet* package (Simon et al., 2011) with cross-validated lambda values. Almost identical results are also derived with Balanced Random Forest approach (estimated in *randomForest* package in Liaw and Wiener 2002). The seven variables have the highest Mean decrease Gini estimates in a random forest model after 2,500 trees and with 9 variables randomly sampled for each tree. There is again a substantial decrease in estimates after the seven values. Mean decrease accuracy gives similar ranking, with the exception that z_5 is the 8th highest predictor and z_{74} the 18th highest. The accuracy of this estimate is highly variable with the choice of number of variables sample and likely not as accurate (possibly due to the fact that error rate for the presence of [s] group is high in the model — 74.2%). The value of variables were chosen based on smallest OOB error rate (tried on a range from 9 to 15 with 2,500 trees). We sample 271 variables from each group (presence vs. absence of [s]) each time to correct for the unbalanced sample.

Several methods for finding the features that predict presence or absence of [s] are thus used. Logistic regression is presented here because it is the simplest and easiest to interpret. In future work, a combination of techniques is recommended to be used for exploratory purposes in a similar way as proposed in this paper.

To conduct an independent generative test of whether the chosen values correlate with [s] in the output data of the Generator network, we set values of the seven identified predictors (z_5, z_{11}, z_{49}, z_{29}, z_{74}, z_{26}, z_{14}) to the marginal value of 1 or -1 (depending on whether the correlation is positive

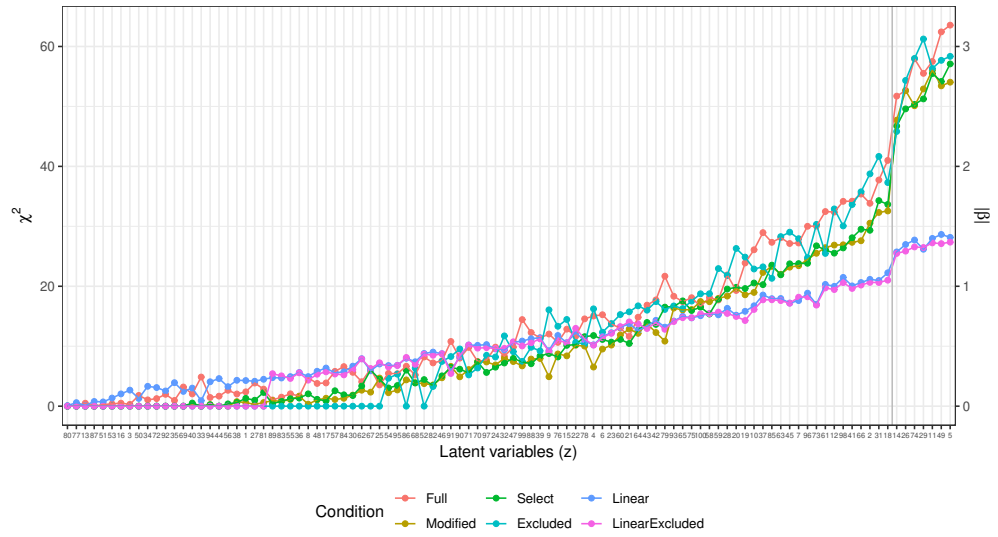

Figure 14: Plot of χ^2 values (left scale) for the 100 predictors across the four generalized additive models. For the two linear models (LINEAR and LINEAR EXCLUDED), estimates of slopes in absolute values ($|\beta|$) are plotted (right scale). The blue vertical line indicates the division between the seven chosen predictors and the rest of the predictor space with a clear drop in estimates between the first seven values (z_5, z_{11}, z_{49}, z_{29}, z_{74}, z_{26}, z_{14}) and the rest of the space.

or negative; see Figure 15) and generated 100 outputs. Altogether seven values in the latent space were thus manipulated, which represents only 7% of the entire latent space. Of the 100 outputs with manipulated values, 73 outputs included a [s] or [s]-like element, either with the stop closure and vowel or without them. The rate of outputs that contain [s] is thus significantly higher when the seven values are manipulated to the marginal levels compared to randomly chosen latent space. In the output data without manipulated values, only 271 out of 3800 generated outputs (or 7.13%) contained an [s]. The difference is significant ($\chi^2(1) = 559.0, p < 0.00001$).

High proportions of [s] in the output can be achieved with manipulation of single latent variables, but the values need to be highly marginal, i.e. extend well beyond the training space. Setting the z_{11} value outside the training range to -15, for example, causes the Generator to output [s] in 87 out of 100 generated (87%) sequences, which is again significantly more than with random input ($\chi^2(1) = 792.7, p < 0.0001$). With value of z_{11} at -25, the rate goes up to 96 out of 100, also significantly different from random inputs ($\chi^2(1) = 959.8, p < 0.0001$).

To further confirm that the regression models identify the variables involved with the presence of [s] in generated outputs, another generative experiment was conducted. We set values of the seven as well as 24 other latent variables to a marginal level well beyond the training range, to ± 4.5, generate samples, and analyze the outputs. Values of 31 latent variables z were thus manipulated, 25 with the highest estimates based on regression models in Figure 14. After the variable with the 25th highest estimate, we picked 6 additional variables that are distanced from the 25th variable in increments of 5. All other variables are sampled randomly and held constant across all the samples, with the exception of the variable in question at a time that is set to ± 4.5. 31% of the latent variables were thus manipulated. One hundred outputs are generated for each analyzed latent variable. Altogether 31×100 (3,100) outputs were thus analyzed and marked for the presence or absence of [s] in the output. Variable z_{14} is excluded from the count, because the output contains frication noise that falls between [s] and [s]-like aspiration, which were difficult to classify (also, the

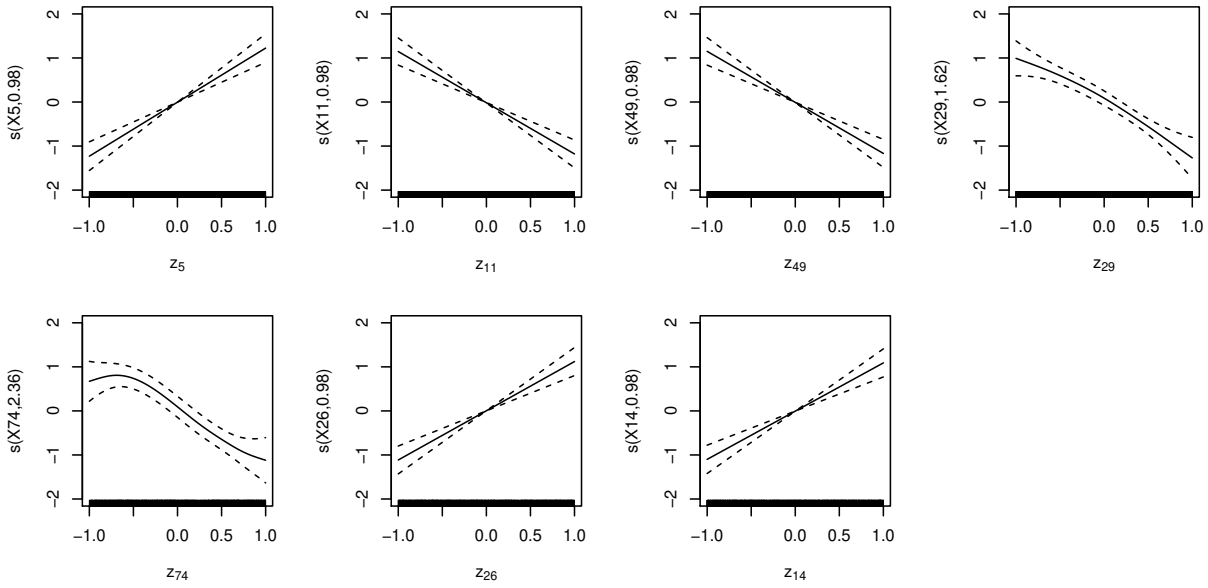

Figure 15: Plots of seven smooth terms with highest χ^2 values in a generalized additive logistic regression model with all 100 latent variables (z) as predictors, estimated with penalty for each term (SELECT). Many of the predictors show linear correlation, which is why a linear logistic regression outputs similar estimates.

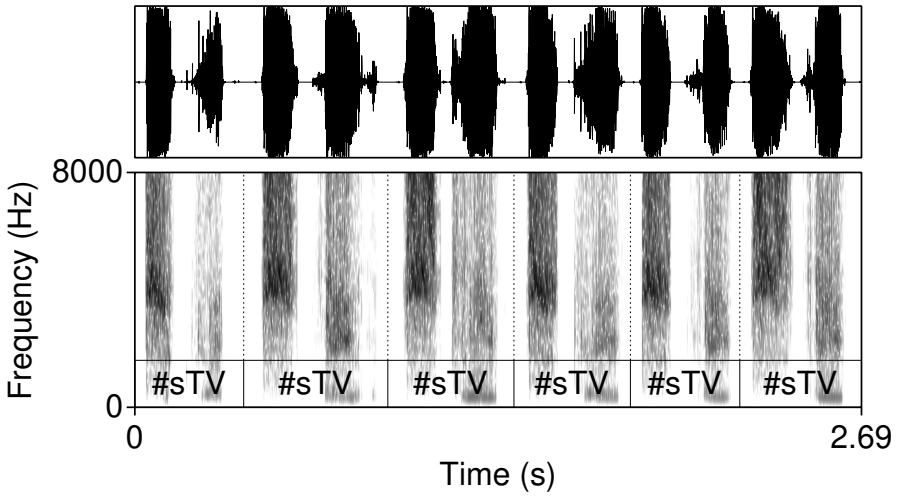

Figure 16: Seven waveforms and spectrograms (0-8000 Hz) of outputs of the Generator network trained after 12,255 steps with the value of z_{11} set at -25. In 96 out of 100 generated samples, the network outputs a sequence containing an [s]. With such a low value of z_{11} (that correlates with amplitude of frication noise), the amplitude of the frication noise reaches the maximum level of 1 in all 100 generated outputs.

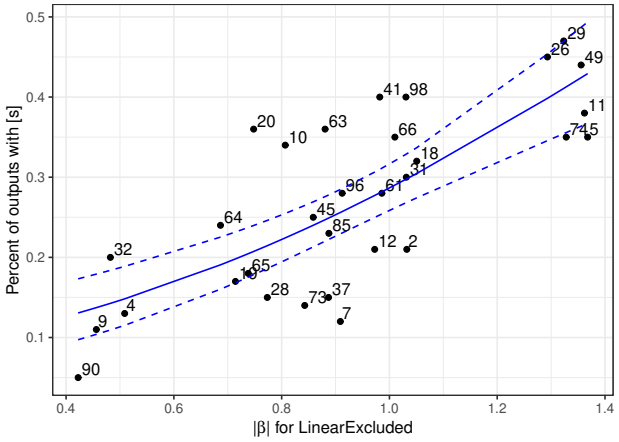

Figure 17: Plot of absolute values of estimates from the LinearExcluded model for 31 analyzed latent variables z (numbered on the plot) and the percent of outputs that contain an [s] based on 100 generated samples. Blue solid line represents predicted values based on the beta regression model with estimates of the LinearExcluded model as the predictor; the dashed lines represent 95% confidence intervals.

target for [s]-like outputs in this variable is closer to 2.5).

The proportion of output containing an [s] (out of 100 generated samples for each of the 31 variables) was fit to a beta regression linear model[7] (using *mgcv* package; Wood 2011). To test which model in Figure 14 is best at predicting the rates of [s] in the outcome, six models were fit with estimates from the six regression models (Full, Select, Modified, Excluded, Linear, LinearExcluded) as the independent variable (Table 14). The best-fitting model was chosen based on AIC: estimates of z-variables in the LinearExcluded model (Figure 14) make the best predictions regarding the presence of absence of [s] in the output.[8] There exists a significant linear correlation between the estimates of the regression models and the actual proportion of generated outputs with [s]: $\beta = 1.70, z = 6.09, p < 0.0001$. In other words, the technique for identifying latent variables that correlate with presence of [s] in the output based on regression models (in Figure 14) successfully identifies such variables. This is confirmed independently: the proportion of generated outputs containing an [s] for each variable z correlates significantly with its estimates from the regression models.

Manipulating some latent variables to the marginal level of ± 4.5 causes a high proportion of [s] in the output despite the regression model estimating their contribution lower than the seven latent variables, as is clear from Figure 17. The two variables with the highest proportion of [s] in the output that are estimated substantially lower than the seven variables are especially z_{41} and z_{98}. There is a clear explanation for the discrepancy of the regression estimates and the rates of [s]-outputs for such variables. While outputs at the marginal values of of the two variables (at ± 4.5) do indeed contain a high proportion of [s]-outputs, the frication noise ceases during the $(-1, 1)$ interval on which the model is trained. Because the regression model only sees the training interval $(-1, 1)$ (annotations are performed on this interval) and does not access outputs with variables outside of this interval, the estimates are consequently lower than the outputs at the marginal levels. There is only a handful of such variables, and since we are primarily interested

[7] Generalized additive models do not provide a better fit and in none of the six models is a smooth significantly different from a linear line.

[8] LinearExcluded is also the model with best AIC score (Table 4).

in those variables that correspond to [s] both within the training range and outside of it, we focus our analysis below to the seven variables in 3.5. The problem with variables in which [s] outputs are present predominantly outside of the training range is the possibility that the [s]-output in these types of cases is secondary/conditioned on some other distribution, because it was likely not encoded in the training stage.

While there is a consistent drop in estimates of the regression models after the seven identified variables (Figure 14) and while several independent generation tests confirm that the seven variables correspond to presence of [s] in the output, the cutoff point between the seven variables and the rest of the latent space is still somewhat arbitrary. It is likely that other latent variables directly or indirectly influence the presence of [s] as well: the learning at this point is not yet categorical and several dependencies not discovered here likely affect the results. Nevertheless, further explorations of the latent space suggest the variables identified with the logistic regression (and other) models (Table 14) are indeed the main variables involved with the presence or absence of [s] in the output.

Additionally, if at the value of z that so substantially exceeds the training range (± 4.5) the latent variable does not influence the outcomes substantially and only marginally increases the proportion of [s]-outputs, as is the case for the majority of the latent variables outside of the seven chosen ones, it is likely that its correlation with [s] in the output is secondary and that the variable does not contribute crucially to the presence of [s].

3.5 Interpolation and phonetic features

Fitting the annotated data and corresponding latent variables from the Generator network to generalized additive and linear logistic regression models identifies values in the latent space that correspond to presence of [s] in the output. As will be shown, below, this is not where exploration of Generator's internal representations should end. We explore whether the mapping between the uniformly distributed input (z) variables that the Generator learns to map to output signal that resembles speech can be associated with specific phonetic or phonological features in that output. The crucial step in this direction is to explore values of the latent space with phonetic correlates in the output beyond the training range, i.e. beyond $(-1, 1)$. Crucially, we observe that the Generator network, while being trained on latent space limited to the range $(-1, 1)$, learns representations that extend this range. Even if the input latent variables (z) exceed the training range, the Generator network outputs samples that closely resemble human speech. Furthermore, the dependencies learned during training extend outside of the $(-1, 1)$ range. Exploring phonetic properties at these marginal values might reveal the actual underlying function of each latent variable.

To explore phonetic correlates of the seven latent variables, we set each of the seven variables separately to the marginal value -4.5 and interpolate to its opposite marginal value 4.5 in 0.5 increments, while keeping randomly-sampled values of the other 99 latent variables z constant. Seven sets of generated samples are thus created, one for each of the seven z values (with the other 99 z-values randomly sampled, but kept constant for all seven manipulated variables). Each set contains a subset of 19 generated outputs that correspond to the interpolated variables from -4.5 to 4.5 in 0.5 increments (again with the constant value of the other 99 z-values). Twenty-nine such sets containing an [s] in at least one set are extracted for analysis.

A clear pattern emerges in the generated data: the latent variables identified as corresponding to the presence of [s] via regression (Figure 14) have direct phonetic correlates and cause changes in amplitude and presence/absence of frication noise of [s] when each of the seven values in the latent space are manipulated to the chosen values, including values that exceed the training range. In other words, by manipulating the identified latent variables, we control the presence/absence of [s] in the output as well as the amplitude of its frication noise.

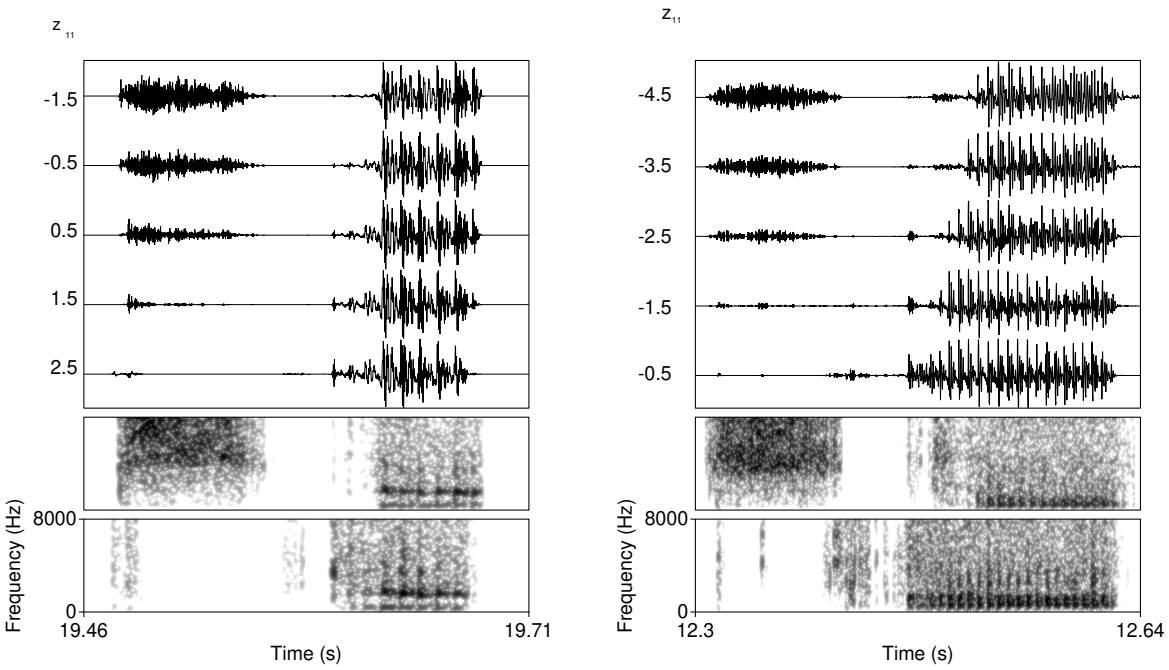

Figure 18: Waveforms and two spectrograms (both $0-8,000$ Hz) of generated data with z_{11} variable manipulated and interpolated. The values on the left of waveforms indicate the value of z_{11}. The two spectrograms represent the highest and the lowest value of z_{11}. A clear attenuation of the frication noise is visible until complete disappearance.

Figure 18 illustrates this effect. Frication noise of [s] gradually decreases by increasing the value of z_{11} until it completely disappears. The exact value of z_{11} for which the [s] disappears differs across examples and likely interacts with other features. It is possible that frication noise in the training has a higher amplitude in some conditions, which is why such cases require a higher magnitude of manipulation of z_{11}. The figure also shows that as the frication noise of [s] disappears, aspiration of a stop in what appears to be a #TV sequences starts surfacing and replaces the frication noise of [s]. Occasionally, frication noise of [s] gradually transforms into aspiration noise. The exact transformation is likely dependent on the 99 other z-variables held constant and their underlying phonetic effect. Regardless of the underlying phonetic effect of the other variables in the latent space, we can force [s] in the output when generating data and manipulating the chosen variables.

To test the significance of the effects of the seven identified features on the presence of [s] and the amplitude of its frication noise, the 29 generated sets of 19 outputs (with z-value from -4.5 to 4.5) for each of the seven variables were analyzed. The outputs were manually annotated for frication noise of [s], closure, VOT, and the following vowel. Outputs gradually change from #sTV to #TV. Only sequences containing an [s] were analyzed; when an output was analyzed as not containing an [s], annotations were stopped and the outputs were not further analyzed. Altogether 161 trajectories were thus annotated; the total number of data points measured is 1,088 because each trajectory contains a few measurements of the interpolated values of z. For each datapoint, maximum intensity of the fricative and the vowel was extracted in Praat (Boersma and Weenink, 2015) with 13.3ms window length (with parabolic interpolation).[9] Figure 19 illustrates how manipulating the values of z of the chosen variables from the marginal value -4.5 decreases

[9]The script used for this task was Lennes (2003).

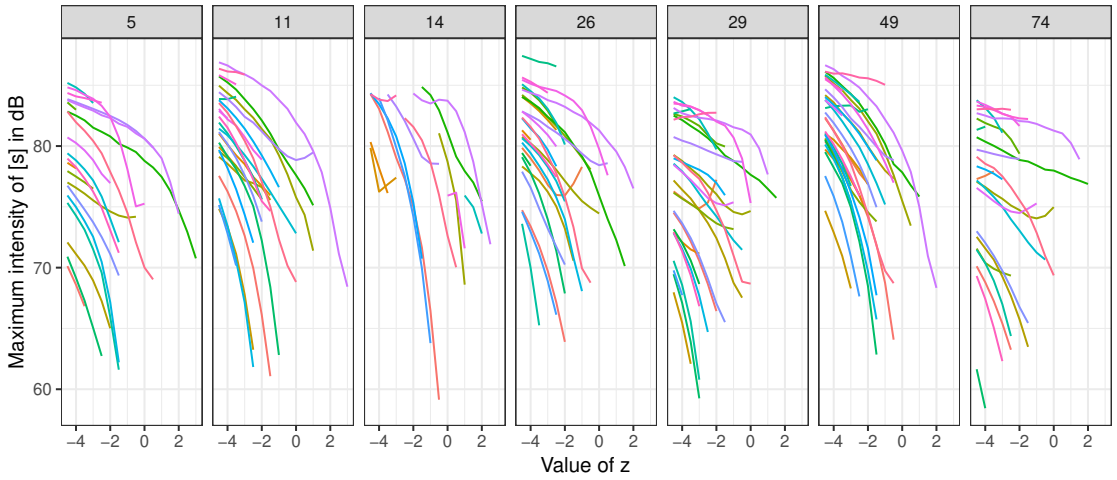

Figure 19: Plots of maximum intensity (in dB) of the fricative part in #sTV sequences when values of the seven z-variables are interpolated from the marginal values ±4.5 in 0.5 increments. Each set of generated samples with the randomly sampled latent variables held constant is colored with the same color across the seven z-variables.

frication noise in the output until [s] in completely absent.

To test whether the decreased frication noise is not part of a general effect of decreased amplitude, we perform significance test on the ratio of maximum intensity between the frication noise of [s] and the following vowel in the #sTV sequences. Figure 20 plots the ratio of maximum intensity of the fricative divided by the sum of two maximum intensities: of the fricative ([s]) and of the vowel (V). The manipulated z-values are additionally normalized to interval [0,1], where 0 represents the most marginal value with [s] (usually ±4.5; referred to as STRONG henceforth) and 1 represents the last value before [s] disappears (WEAK). Note that the point at which [s] is not present in the output anymore, but the vowel still surfaces (which would yield the ratio at 0) is not included in the model.

The data were fit to a beta regression generalized additive mixed model (in *mgcv* package; Wood 2011) with the ratio as the dependent variable, the seven chosen variables as the parametric term, thin-plate smooths for each variable and random smooths (with first order of penalty; Baayen et al. 2016; Sóskuthy 2017) for (i) trajectory and for (ii) value of other variables in the latent space of the Generator network. Number of knots is chosen as default in the smooth term and as 5 in the random smooths. There is negative autocorrelation at lag 1, but with so little variance left unexplained (99.5%; adjusted $R^2 = 0.99$), this likely does not affect outcomes substantially (Sóskuthy, 2017).[10] Figure 20 plots the normalized trajectories of the ratio and predicted values based on the generalized additive model. All smooths (except for z_{74}) are significantly different from 0 (all coefficients in Table 7) and the plots show a clear negative trajectory. In other words, maximum intensity of [s] is increasingly attenuated compared the the intensity of the vowel as z approaches the opposite value from the one identified as predicting the presence of [s] until it completely disappears from the output.

The seven variables thus strongly correspond to presence or absence of [s] in the output; by

[10] Autocorrelation is reduced when the ratio is modeled as normally distributed and correction for AR(1) correlation is added to the model with $\rho = 0.98$ (Baayen et al., 2016). This, however, introduces a substantially worse fit. Since estimates of the smoothing terms are similar (with the same smooths being significant), we keep the beta regression model with autocorrelation.

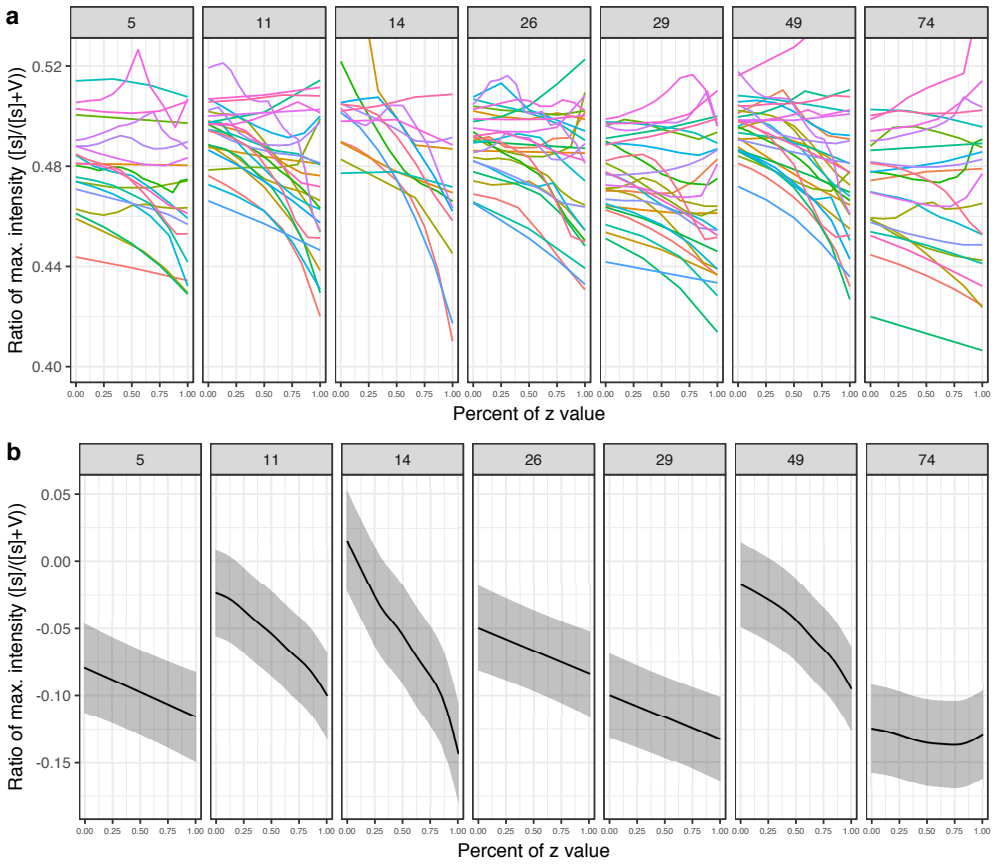

Figure 20: **(a)** Plots of ratios of maximum intensity between the frication of [s] and phonation of the vowel in #sTV sequences across the seven variables. The interpolated values are normalized where 0 represents the most marginal value of z with [s] in the output and 1 represents the value of z right before [s] ceases from the output. Four marginal values are left out from the plot (but are included in the models). Each set of generated samples with the randomly sampled latent variables held constant is colored with the same color across the seven z-variables. **(b)** Predicted values with 95% CIs of the ratio based on beta regression generalized additive model (Table 7) across the several variables with normalized values.

manipulating the chosen variables we can attenuate frication noise of [s] and cause its complete disappearance in the generated data. Again, the discovery of these features is possible because we extend the initial training range and test predictions on marginal values.

3.6 Phonetic values of latent variables

Interpolation of latent variables reveals that presence of [s] is not controlled by a single latent variable, but by at least seven of them. Additionally, there appears to be no categorical cut-off point in the magnitude of the effect between the variables, only a steep drop of regression estimates and a decline of outputs with [s] in generated data. This suggest that the learning at this stage is gradient and probabilistic rather than fully categorical.

The different latent variables that correspond to presence of [s], however, are not phonetically vacuous: individually, they have distinct phonetic correspondences. The generated samples reveal that the variables' secondary effect (besides outputting [s] and controlling its intensity) is likely reflected in spectral properties of the frication noise. The seven variables are thus similar in the sense that manipulation of their values results in presence of [s] by controlling its frication noise. They crucially differ, however, in the effects on the spectral properties of the outputs.

To test this prediction, spectral properties of the output fricatives are analyzed. The same 29 sets of generated samples are used in the analysis; one z-value is manipulated in each set while other variables are sampled randomly and held constant. The marginal values of the variables were chosen for this test: the values with the strongest presence of [s] (which in most cases is ± 4.5; henceforth STRONG) and the value before which [s] ceases from the output (henceforth WEAK). Center of gravity (COG), kurtosis, and skew of the frication noise was analyzed with a script in (Rentz, 2017) in Praat (Boersma and Weenink, 2015). Period of frication is sliced into 10% intervals. The data includes 161 trajectories (from the 29 generated sets) and $161 \times 10 = 1,610$ unique datapoints. COG, kurtosis, and skew based on power spectra are measured in each of these 1,610 intervals with 750–8,000 Hz Hand band pass filter (100 Hz smoothing). Results were fit to six generalized additive mixed models with COG, kurtosis, and skew as the dependent variables (3 for each of the levels STRONG and WEAK). The parametric terms included the seven latent variables z. The smoothing terms included smooths for latent variable z_{11} and difference smooths for the other six variables z. The model also includes random smooths for each fricative (from 10 to 100% with 10 knots) and for each of the 29 generated sets with equal random values of other 99 z-variables (with 7 knots; random smooths are fitted with first order of penalty, see Baayen et al. 2016; Sóskuthy 2017). The models were fit with correction for autocorrelation with ρ-values ranging from 0.15 to 0.7.

Spectral properties of the generated fricatives are generally not significantly different at the value of z right before [s] disappears from the outputs (WEAK; left column in Figure 21). As values of z increase towards the marginal levels (in most cases, ± 4.5), however, clear differentiation in spectral properties emerge between the seven z-variables (STRONG; right column in Figure 21). The trajectory for center of gravity, for example, significantly differs between z_{11} and most of the other six variables. Overall kurtosis is significantly different when z_{11} is manipulated, compared to, for example, z_{26} and z_{29}. Similarly, while z_{74} does not significantly attenuate amplitude of [s], it does significantly differs in skew trajectory of [s]. The main function of z_{74} is thus likely in its control of spectral properties of frication of [s] (e.g. skew). For all coefficients and significant relationship of the six models, see Tables 8, 9, 10, 11, 12, and 13.

In sum, manipulating the latent variables that correspond to [s] in the output not only attenuates frication noise independent of the vocalic amplitude and causes [s] to surface or disappear from the output, but the different z-variables likely correspond to different phonetic features of the frication

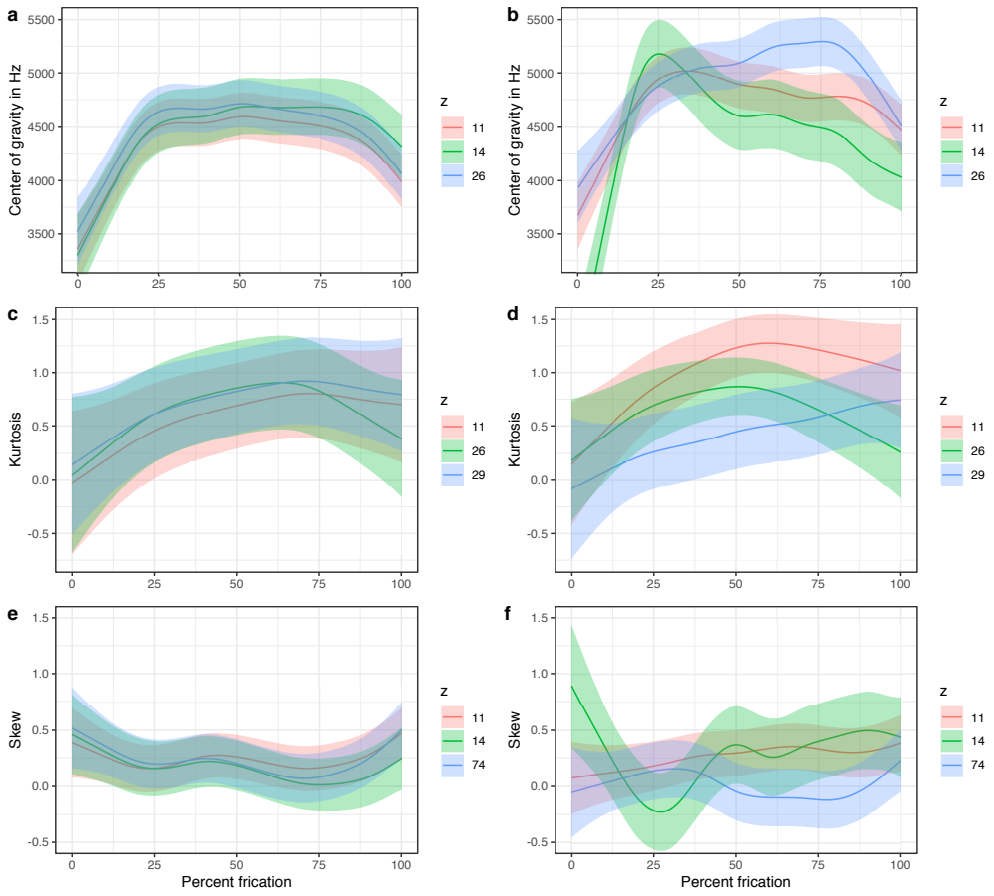

Figure 21: Predicted values of COG, kurtosis, and skew with 95% CIs in two conditions: WEAK with z-variables at the value before [s] ceases from the output (right column) and STRONG with the most marginal value with [s]-output (± 4.5 in most cases). Predicted values are based on generalized additive models in Tables 8, 9, 10, 11, 12, and 13. The plots show a clear differentiation from no significant differences in COG, kurtosis, and skew, to clear significant overall differences and trajectory differences as the z-values move from WEAK toward the marginal (STRONG) values. Difference smooths for the presented variables are in Figure 22.

noise. At the level before the frication noise ceases from the output, the are no differences in spectral moments between the latent variables. By setting the values to the marginal levels well beyond the training range, however, significant differences emerge both in overall levels as well as in trajectories of COG, kurtosis, and skew. It is thus likely that the variables collectively control the presence or absence of [s], but that individually, they control various phonetic features — spectral properties of the frication noise.

4 Discussion

The Generator network trained after 12,255 steps learns to generate outputs that closely resemble human speech in the training data. The results of the experiment in Section 3.3 suggest that the generated outputs from the Generator network replicate the conditional distribution of VOT duration in the training data. The Generator network thus not only learns to output signal that

resembles human speech from noise (input variables sampled from a uniform distribution), but also learns to output shorter VOT durations when [s] is present in the signal. While this distribution is phonologically local, it is non-local in phonetic terms as a period of closure necessarily intervenes between [s] and VOT.

4.1 Parallels in human behavior

While the generated outputs contain evidence that the network learns the conditional distribution of VOT duration, some outputs still violate this distribution. In fact, the Generator occasionally outputs VOT durations in the #sTV condition that are longer than all VOT durations in training data in the same condition. This suggests that the model does not categorically learn the conditional distribution yet and will occasionally violate it. These outputs resemble human behavior in L1 acquisition. Infants acquiring VOT in English undergo a period in which they produce VOT durations substantially longer compared to the adult input, not only categorically in all stops (Macken and Barton, 1980; Catts and Jensen, 1983; Lowenstein and Nittrouer, 2008), but also in the position after the sibilant [s]. McLeod et al. (1996) studied acquisition of #sTV and #TV sequences in 2;0 to 2;11 year old children. Unlike the Generator network, children often simplify the initial clusters from #sTV to a single stop #TV. What is parallel to the outputs of the Generator, however, is that the VOT duration of the simplified stop is overall significantly shorter in underlying #sTV sequences, but there exist a substantial period of variation and occasionally the language-acquiring children output long-lag VOT durations there (McLeod et al. 1996, for similar results in language-delayed children, see Bond 1981). Bond and Wilson (1980) present a similar study, but include older children that do not simplify the #sT cluster. This group behaves exactly parallel to the Generator's network: the overall duration of VOT in the #sTV sequences is shorter compared to the #TV sequences, but the longest duration of any VOT is attested once in the #sTV, not in the #TV condition (Bond and Wilson, 1980). The children thus learn both to articulate the full #sT cluster and to output a shorter VOT durations in the cluster condition. Occasionally, however, they output a long-lag VOT in the #sTV condition that is longest than any VOT in the #TV condition. The purpose of these comparisons is not to suggest the GAN model learns the data in exactly the same manner as human infants, but to suggest that clear similarities exist in behavior between the proposed model and human behavior in speech acquisition.

Further parallels to the Generator's behavior are available in L2 acquisition, speech errors, and speech impairment. Results from studies on L2 acquisition of the aspiration contrast in #sTV and #TV sequences suggest that learners start with a smaller distinction between the two groups and acquire the non-aspiration after [s] after more exposure (Haraguchi, 2003). A smaller initial difference between the two conditions in L2 acquisition, for example, improves from a group with little exposure to English to a group with more exposure in Japanese learners (Haraguchi, 2003). Saudi Arabic L2 learners of English produce substantially longer VOT durations in #sTV sequences compared to the native inputs (Alanazi, 2018), which resembles imperfect learning in the Generator's network. Speech errors also provide a parallel to the described behavior of the Generator network. German has a similar process of aspiration distribution as English. In an experiment of elicited speech errors, German speakers produced aspirated stops with longer VOT durations in erroneous sequences with inserted sibilant in 34% of cases (Pouplier et al., 2014). This suggests that the allophonic rule fails to apply in the speech errors. Similarly, the Generator fails to output unaspirated stops after a sibilant [s] in a subset of examples. Finally, Buchwald and Miozzo (2012) analyzed VOT durations of two patients with cluster production errors. One patient outputs long VOT durations in the #sTV condition (after the cluster simplified) that correspond to VOT durations of the singleton stops. The other patient's VOT duration correctly follows the distribution

between the two conditions (#TV vs. #sTV) even when the cluster is simplified. It is hypothesized that the first patient with long VOT durations in the #sTV condition exhibits traces of impairment of phonological computation, while the second patient shows traces of phonetic execution impairment. The Generator models fails to categorically output shorter VOT duration, which would suggest imperfect phonological computation, as predicted by the fact that the articulatory component is absent from our model (see also discussion below).

The network also generates segmentally innovative outputs for which no evidence was available in the training data. Producing novel sequences with only the sibilant [s] and the vowel without the intervening consonant (#sV) or with two stops (#TTV) suggests that the network treats the period of frication [s] or the period of closure, burst, and aspiration (of a stop) as units that can be recombined with other units into innovative outputs. That segments are learned as units that can be manipulated and recombined is additionally suggested by exploration of the Generator's latent variables (Section 3.5).

A subset of the innovative outputs is consistent with linguistic behavior in humans. The Generator's innovative outputs thus closely resemble one of the main traits of human phonology: productivity. Human subjects are able to evaluate and produce nonce-words even if a string of phonemes violates language-specific phonotactics, as long as the basic universal phonotactic requirements that treats phones as atomic units are satisfied (for an overview of phonotactic judgments, see Ernestus 2011 and literature therein). Deleting or inserting segments are also common patterns in both L1 acquisition (Macken and Ferguson, 1981), loanword phonology (Yildiz, 2005), in children with speech disorders (Catts and Kamhi, 1984; Barlow, 2001), as well as in speech errors (Alderete and Tupper, 2018b). For example, #sT clusters are often simplified in L1 acquisition (Gerlach, 2010). While the most common outcome is deletion of [s], deletion of the stop is robustly attested as well in L1 acquisition, both in the general population and in infants with speech disorders (Catts and Kamhi, 1984; Ohala, 1999; Gerlach, 2010; Syrika et al., 2011). While the reduction in L1 acquisition likely involves articulatory factors that are lacking in our model, the fact that segmental units can be dropped and recombined in L1 acquisition resembles the Generator's innovative outputs — #sV sequences.

These innovative outputs of the Generator's network have potential for contributing to our understanding of evolution of phonology as well (for an overview of the field, see Gibson et al. 2012). The main process that any model of the evolution of phonology needs to explain is the change from "holistic" acoustic signals in the proto-language to the "combinatorial" principle that operates with discrete units — phonemes and their combinations (Oudeyer, 2001, 2002, 2005, 2006; Zuidema and de Boer, 2009). The Generator network shows precisely this behavior: in addition to learning to reproduce the input, it learns to treat some phonetic content as units that can be recombined into novel and unobserved sequences. The fact that the networks attempts to recombine segments as units to novel unobserved sequences bears the potential for explaining how segmentation in human phonology emerges with neural architecture only. Atomic lexicalized items at the proto-language stage can automatically develop into a segmented string of units using only the mechanisms we observe in the proposed model. The advantage of the GAN model over competing proposals (Oudeyer, 2001, 2002, 2005, 2006; Zuidema and de Boer, 2009) is that learning is completely unsupervised and that the network's only input are raw acoustic data. The exact details of modeling phonological evolution with Generative Adversarial architecture is beyond the scope of the present paper.

4.2 Latent variables as correlates of features

Finally, the paper proposes a technique for recovering internal representations of the Generator network. The first crucial observation is that the dependencies learned in the latent space limited by some interval extend beyond that interval in what appears to be a linear relationship. This allows for an in-depth analysis of phonetic effects of each latent variable in the generated data. Non-linear regression of annotated data and the latent variables identifies those variables that strongly correlate with the presence of [s] in the output. Manipulating values of the identified latent variables, both within the training interval and outside of it, results in significantly higher rates of [s] in the output. By interpolating values of individual latent variables outside of the training interval, we explore the exact phonetic correlates of each latent variable. The results suggest that the Generator network learns to use latent variables to encode imperfect equivalents of phonetic features. Since the features not only correspond to phonetic properties, but to the categorical presence or absence of [s] in the output, the network not only uses latent space to encode phonetic features, but also what would be an approximate equivalent of phonological features — absence or presence of a segment. What is unique about the network is that phonetics and phonology are modeled simultaneously.

While the presence of [s] in the output is controlled by multiple latent variables, each of the variables has an underlying phonetic function. While there are no significant differences in phonetic correlates of z-variables when their value is at the last point before ceasing from the output, a clear differentiation emerges when the values are set to the marginal level (Figure 21). Each of the seven variables thus has a clear phonetic function: the seven variables at the marginal levels differ significantly in spectral moments of frication noise both in overall values or in trajectories along the fricative duration.

Features have long been in the center of phonetic and phonological literature (Trubetzkoy, 1939; Chomsky and Halle, 1968; Clements, 1985; Dresher, 2015; Shain and Elsner, 2019). Extracting features based on unsupervised learning of pre-segmented phones using autoencoder neural networks has recently seen success. Shain and Elsner (2019) train an autoencoder with binary stochastic neurons on pre-segmented speech data and argue that bits in the code of the autoencoder network correspond to phonological features as posited by phonological theory. As was argued in Section 3.4, our model shows traces of imperfect self-organizing of phonetic (e.g. spectral moments) and phonological (e.g. presence of [s]) features in the latent space, while learning allophonic distributions at the same time. Considerable differences between the theoretically assumed features and our results, of course, remain. Latent space encoding in our model resembles phonological features or feature matrices (such as full presence of [s] in the output) and phonetic features (such as COG or kurtosis), but the relationships are gradient and not categorical. The current model also does not test whether higher order grouping of phonemes in accordance with some abstract feature such as [+sonorant] emerge in the training. This task is left for future work. Despite these differences, the fact that we can actively control presence of [s] and its spectral properties in the generated data with a subset of latent variables suggest that the network learns to encode information in its latent space that resembles phonetic and phonological features.

On a very speculative level, the latent space of the Generator's network might have an approximate correlation in featural representation of speech production in human brain, where featural representations are also gradient and involve multiple correlates. Bashivan et al. (2019) argue for the existence of direct correlations between the neural network architecture and vision in human brain. Similarly, Guenther and Vladusich (2012), Guenther (2016), and Oudeyer (2005) propose models of simple neural maps that might have direct equivalents in neural computation of speech planing with some actual clinical applications that result from such models. Recently, high-density direct cortical surface (electrocorticographic) recordings of superior temporal gyrus during open

brain surgery in Mesgarani et al. (2014) suggests that recorded brain activity has direct correlates in encoding of phonetic features. Encoding for phonetic and phonological features in the latent space of the Generator's network can speculatively be compared to such brain recordings that serve as the basis for articulatory execution. The correspondences between the brain activity and phonetic and phonological features are multiple and gradual, not categorical, which bear resemblances to our model. To be sure, this comparison can only be indirect and speculative at this point.

4.3 Improvements

Among the objections against modeling phonological learning with Generative Adversarial Networks might be that the model is too powerful and overgenerates. First, it has been shown on numerous examples that phonology, while being computationally limited (Heinz, 2011; Avcu et al., 2017), is more powerful than the attested phonological typology. For example, alternations that never surface in natural languages are learned in the artificial grammar learning paradigm (Glewwe, 2017; Glewwe et al., 2017; Beguš, 2018a,b; Avcu, 2018). Secondly, overgeneration is a less severe violation than undergeneration. Absence of unnattested patterns that are derivable within a theory can be explained with external factors, such as historical developments or articulatory limitations. Not generating attested patterns, however, is more serious: a model of phonology should at minimum derive the observed phonological processes. Finally, the main reason the model overgenerates is because the current proposal involves no information about the articulatory mechanism in speech production. In other words, the GAN model is completely unconstrained for articulatory mechanisms. This fact would pose a problem if the current model's claim were that the network models phonetic and phonological in their entirety. The aims of the current proposal, however, are more restricted. The network models learning without any articulatory information. Lack of articulatory information in the model (and consequently, the overgeneration problem) might in fact be an advantage of the current proposal. Learning in phonetics and phonology is often treated as if it is homogeneous, involving a single mechanism. It is likely, however, that learning in phonetics and phonology involves various different levels and mechanisms. Motor-planing learning on the articulatory level is likely different from learning of articulatory targets based on perception or of abstract symbol manipulation on the phonological level, even though acquisition of one feeds into the other. The best evidence that phonetic and phonological acquisition involves different levels are aphasia patients with different production errors Buchwald and Miozzo (2012). If impairment targets the motor-planning unit, the phonological level is intact and the production error causes only deletion of [s] in #sTV target clusters with the stop being unaspirated, as predicted by phonology. If, on the other hand, phonological computation is impaired, the stop surfaces as aspirated, similar to the outputs of our GAN model. By excluding articulatory information, we model phonetic and phonological learning as if they were unconstrained by articulators, but only by cognition. In other words, we model phonological computation on a cognitive level as if no articulatory constraints were present in human speech. This is highly desired for the task of distinguishing those aspects of phonology that are influenced by cognitive factors from those that are influenced by articulation, motor planning, or historical developments (Beguš, 2018a).

Several further explorations and improvements of the model are warranted. The acoustic speech data fed to the network is modeled as waveform data points, i.e. pressure points in a time continuum (as proposed for WaveGAN in Donahue et al. 2019). This has considerable advantages for exploring the properties of the network, because spectral analysis introduces losses in the signal. A GAN that would be trained on spectral transformations would likely be closer to reality, as human auditory mechanisms resembles spectral information more closely than raw pressure points (Young, 2008; Pasley et al., 2012; Mesgarani et al., 2014). Adding an articulatory model would likewise yield

novel information on the role of articulatory learning on phonetic and phonological computation.

5 Conclusion and future directions

The results of this paper suggest that we can model phonology not only with rules (as in rule-based approaches), input-output optimization (as in Optimality Theory), or with neural network architecture that already assumes some level of abstraction, but with complex neural networks that are trained in an unsupervised manner from raw phonetic data. The Generative Adversarial model of phonology (trained on an implementation of DCGAN architecture for audio data in Donahue et al. 2019) derives outputs that resemble speech from latent variables. The network learns to encode phonetic and phonological information in the latent space by a process that has a direct parallel in human speech acquisition: imitation. The results of the computational experiment suggest that the network learns conditional allophonic distribution of VOT duration. We propose a technique that identifies variables that correspond to presence of [s] in the output and show that by manipulating these values, we can generate data with or without [s] in the output as well as control its intensity as well as spectral properties of its frication noise. While at least seven latent variables control presence of [s], each of them has a clear phonetic function. Finally, the model generates innovative outputs, suggesting its productive nature and the ability to treat phonetic elements as units that can be productively recombined. The behavior of the model is compared against speech acquisition, speech errors, and speech impairment; several similarities are identified.

The current proposal models one allophonic distribution in English. Training GAN networks on further processes and on languages other than English should yield more information about learning representations of different features, phonetic and phonological processes, and about computational models of the cognitive aspects of human speech production and perception in general. The paper outlines methodology for establishing internal representations and testing predictions against generated data, but represents just a first step in a broader task of establishing learning representation of phonetic and phonological data in a Generative Adversarial framework.

The proposed model also has implications beyond modeling the cognitive basis of human speech. The results of establishing internal representation of the Generator network has implications for more applicable tasks in natural language processing. Identifying latent variables that corresponds to output sounds allows for a model that generates desired input strings with different output properties. Discussing the details of such models is beyond the scope of this paper.

Acknowledgements

This research was funded by a grant to new faculty at the University of Washington. I would like to thank Sameer Arshad for slicing data from the TIMIT database and Heather Morrison for annotating data. All mistakes are my own.

A Supplementary Materials: Regression models

	β	SE	t-value	Pr(>|t|)
(Intercept)	57.4	0.28	203.37	0.0000
#TV vs. #sTV	-32.4	0.95	-34.16	0.0000
[p] vs. mean	-7.8	0.44	-17.64	0.0000
[t] vs. mean	-2.2	0.38	-5.79	0.0000
#sTV:[p]	2.2	1.44	1.49	0.1357
#sTV:[t]	2.8	1.20	2.30	0.0213

Table 5: Coefficients of a linear model with duration of VOT in the training data as the dependent variable and condition (#TV vs. #sTV) and PLACE of articulation (with interaction) as independent variables.

Select				
A. parametric coef.	Estimate	Std. Error	t-value	p-value
(Intercept)	-5.3046	0.2104	-25.2179	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(z_5)	0.9828	9.0	57.0935	0.0000
s(z_{11})	0.9823	9.0	55.4790	0.0000
s(z_{14})	0.9791	9.0	46.7389	0.0000
s(z_{26})	0.9802	9.0	49.5906	0.0000
s(z_{29})	1.6222	9.0	51.2550	0.0000
s(z_{49})	0.9819	9.0	54.1608	0.0000
s(z_{74})	2.3630	9.0	50.3333	0.0000
Linear excluded				
	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-6.1378	0.2879	-21.32	0.0000
z_5	1.3678	0.1770	7.73	0.0000
z_{11}	-1.3619	0.1725	-7.89	0.0000
z_{14}	1.2739	0.1759	7.24	0.0000
z_{26}	1.2932	0.1725	7.50	0.0000
z_{29}	-1.3234	0.1705	-7.76	0.0000
z_{49}	-1.3557	0.1747	-7.76	0.0000
z_{74}	-1.3280	0.1795	-7.40	0.0000

Table 6: Coefficients of the seven predictors with highest χ^2 values or highest slope estimates from two models: SELECT and LINEAR EXCLUDED.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = z_{11}	-0.0571	0.0156	-3.6505	0.0003
z_5	-0.0404	0.0123	-3.2820	0.0011
z_{14}	-0.0011	0.0144	-0.0753	0.9400
z_{26}	-0.0097	0.0115	-0.8444	0.3989
z_{29}	-0.0590	0.0113	-5.2131	< 0.0001
z_{49}	0.0074	0.0112	0.6595	0.5100
z_{74}	-0.0741	0.0121	-6.1071	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(zValuePerc):z_5	1.0002	1.0000	11.8417	0.0006
s(zValuePerc):z_{11}	4.1696	4.8546	14.1190	< 0.0001
s(zValuePerc):z_{14}	5.3322	6.1117	36.6899	< 0.0001
s(zValuePerc):z_{26}	1.0003	1.0002	12.5952	0.0004
s(zValuePerc):z_{29}	1.0002	1.0000	12.0036	0.0006
s(zValuePerc):z_{49}	4.2002	4.8650	19.1225	< 0.0001
s(zValuePerc):z_{74}	3.2768	3.7863	1.2326	0.2479
fs(zValuePerc,sameValues,m=1,k=5)	110.6863	143.0000	6.1542	< 0.0001
fs(zValuePerc,trajectoryZ,m=1,k=5)	558.2060	728.0000	56.9670	< 0.0001

Table 7: Coefficients of a beta regression generalized additive model with ratio of maximum intensity ([s] vs. vowel) as the dependent variable.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = z_{11}	4751.7378	84.7008	56.1002	< 0.0001
z_5	218.6576	116.9490	1.8697	0.0618
z_{14}	-236.4061	134.6301	-1.7560	0.0793
z_{26}	195.2722	108.9736	1.7919	0.0734
z_{29}	103.6866	107.8602	0.9613	0.3366
z_{49}	17.6464	106.4109	0.1658	0.8683
z_{74}	108.7466	113.8531	0.9551	0.3397
B. smooth terms	edf	Ref.df	F-value	p-value
s(zValuePerc) = z_{11}	7.5348	7.9933	12.1238	< 0.0001
s(zValuePerc):z_5	4.5539	5.7457	2.9261	0.0081
s(zValuePerc):z_{14}	7.3604	8.3734	5.7228	< 0.0001
s(zValuePerc):z_{26}	5.5900	6.8683	3.8049	0.0005
s(zValuePerc):z_{29}	5.8536	7.1301	2.9198	0.0045
s(zValuePerc):z_{49}	4.4714	5.6434	1.8590	0.0803
s(zValuePerc):z_{74}	4.2765	5.4186	2.8162	0.0136
fs(zValuePerc,sameValues,m=1,k=10)	143.4989	288.0000	1.0560	< 0.0001
fs(zValuePerc,trajectoryZ,m=1,k=7)	168.3558	1120.0000	0.2032	< 0.0001

Table 8: Coefficients of a generalized additive model with center of gravity as the dependent variable with the marginal value of z-variables (STRONG). The model was fit with correction for autocorrelation with $\rho = 0.7$.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = z_{11}	1.0675	0.1045	10.2167	< 0.0001
z_5	-0.4521	0.1420	-3.1842	0.0015
z_{14}	0.3405	0.1693	2.0105	0.0446
z_{26}	-0.4434	0.1323	-3.3517	0.0008
z_{29}	-0.6225	0.1339	-4.6502	< 0.0001
z_{49}	0.0431	0.1332	0.3234	0.7464
z_{74}	-0.5129	0.1383	-3.7077	0.0002
B. smooth terms	edf	Ref.df	F-value	p-value
s(zValuePerc) = z_{11}	3.3590	4.0455	4.4859	0.0013
s(zValuePerc):z_5	1.0001	1.0001	1.8978	0.1686
s(zValuePerc):z_{14}	5.7086	6.9165	2.9066	0.0054
s(zValuePerc):z_{26}	1.0000	1.0000	2.3717	0.1238
s(zValuePerc):z_{29}	2.2995	2.8348	1.4855	0.2361
s(zValuePerc):z_{49}	5.3523	6.5335	2.5656	0.0106
s(zValuePerc):z_{74}	1.0000	1.0000	0.1912	0.6620
fs(zValuePerc,sameValues,m=1,k=10)	69.5866	288.0000	0.4214	< 0.0001
fs(zValuePerc,trajectoryZ,m=1,k=7)	174.7382	1120.0000	0.2422	< 0.0001

Table 9: Coefficients of a generalized additive model with kurtosis as the dependent variable with the marginal value of z-variables (STRONG). The model was fit with correction for autocorrelation with $\rho = 0.2$.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = z_{11}	0.2726	0.0841	3.2434	0.0012
z_5	-0.2686	0.1197	-2.2448	0.0249
z_{14}	-0.0188	0.1377	-0.1368	0.8912
z_{26}	-0.1965	0.1115	-1.7629	0.0781
z_{29}	-0.2011	0.1101	-1.8270	0.0679
z_{49}	-0.0403	0.1063	-0.3792	0.7046
z_{74}	-0.2468	0.1165	-2.1193	0.0342
B. smooth terms	edf	Ref.df	F-value	p-value
s(zValuePerc) = z_{11}	4.5857	5.4215	1.3591	0.3433
s(zValuePerc):z_5	4.2885	5.5104	2.0917	0.0864
s(zValuePerc):z_{14}	6.4497	7.7372	3.3262	0.0009
s(zValuePerc):z_{26}	6.4653	7.7452	2.2045	0.0303
s(zValuePerc):z_{29}	3.8520	4.9849	2.0158	0.0716
s(zValuePerc):z_{49}	1.0000	1.0001	0.0105	0.9186
s(zValuePerc):z_{74}	4.0239	5.1943	1.9009	0.0916
fs(zValuePerc,sameValues,m=1,k=10)	113.6068	288.0000	0.6943	< 0.0001
fs(zValuePerc,trajectoryZ,m=1,k=7)	0.0001	1120.0000	0.0000	0.9908

Table 10: Coefficients of a generalized additive model with skew as the dependent variable with the marginal value of z-variables (STRONG). The model was fit with correction for autocorrelation with $\rho = 0.7$.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = z_{11}	4396.2895	88.8182	49.4976	< 0.0001
z_5	2.4059	85.3386	0.0282	0.9775
z_{14}	109.3881	101.5196	1.0775	0.2815
z_{26}	98.1943	79.3503	1.2375	0.2162
z_{29}	-34.1064	78.4139	-0.4350	0.6637
z_{49}	-42.5635	77.1872	-0.5514	0.5815
z_{74}	19.5268	85.7248	0.2278	0.8199
B. smooth terms	edf	Ref.df	F-value	p-value
s(zValuePerc) = z_{11}	6.9763	7.4135	16.1815	< 0.0001
s(zValuePerc):z_5	1.0002	1.0003	0.0007	0.9793
s(zValuePerc):z_{14}	2.1327	2.4962	1.0245	0.2793
s(zValuePerc):z_{26}	1.0072	1.0110	0.2178	0.6468
s(zValuePerc):z_{29}	1.0003	1.0004	0.4048	0.5249
s(zValuePerc):z_{49}	1.0001	1.0002	3.5280	0.0606
s(zValuePerc):z_{74}	2.4946	2.9500	1.2008	0.2568
fs(zValuePerc,sameValues,m=1,k=10)	198.6356	288.0000	3.3009	< 0.0001
fs(zValuePerc,trajectoryZ,m=1,k=7)	413.0783	1120.0000	1.6191	< 0.0001

Table 11: Coefficients of a generalized additive model with center of gravity as the dependent variable with the value of z-variables at the point before [s] ceases from the output (WEAK).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = z_{11}	0.6420	0.1544	4.1575	< 0.0001
z_5	-0.0037	0.1463	-0.0256	0.9796
z_{14}	-0.4010	0.1685	-2.3803	0.0174
z_{26}	0.0230	0.1368	0.1678	0.8668
z_{29}	0.0909	0.1344	0.6762	0.4991
z_{49}	0.0870	0.1323	0.6576	0.5109
z_{74}	0.1577	0.1429	1.1032	0.2702
B. smooth terms	edf	Ref.df	F-value	p-value
s(zValuePerc) = z_{11}	2.6481	2.9134	1.5639	0.2107
s(zValuePerc):z_5	1.0000	1.0000	3.2961	0.0697
s(zValuePerc):z_{14}	1.0000	1.0001	0.5569	0.4556
s(zValuePerc):z_{26}	1.9006	2.3489	1.0773	0.3078
s(zValuePerc):z_{29}	1.0000	1.0000	0.0284	0.8661
s(zValuePerc):z_{49}	1.0000	1.0000	0.0002	0.9887
s(zValuePerc):z_{74}	1.3675	1.6177	0.3165	0.5648
fs(zValuePerc,sameValues,m=1,k=10)	181.3987	288.0000	2.3885	< 0.0001
fs(zValuePerc,trajectoryZ,m=1,k=7)	128.9479	1120.0000	0.1673	< 0.0001

Table 12: Coefficients of a generalized additive model with kurtosis as the dependent variable with the value of z-variables at the point before [s] ceases from the output (WEAK). The model was fit with correction for autocorrelation with $\rho = 0.2$.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = z_{11}	0.2432	0.0734	3.3145	0.0009
z_5	-0.0384	0.0758	-0.5067	0.6125
z_{14}	-0.0906	0.0873	-1.0373	0.2998
z_{26}	-0.1433	0.0705	-2.0325	0.0423
z_{29}	-0.0392	0.0698	-0.5613	0.5747
z_{49}	-0.0191	0.0687	-0.2777	0.7813
z_{74}	-0.0151	0.0740	-0.2043	0.8381
B. smooth terms	edf	Ref.df	F-value	p-value
s(zValuePerc) = z_{11}	5.2698	5.9125	3.3712	0.0037
s(zValuePerc):z_5	1.0000	1.0000	0.5871	0.4437
s(zValuePerc):z_{14}	1.0000	1.0000	1.7508	0.1860
s(zValuePerc):z_{26}	1.0000	1.0000	0.1276	0.7210
s(zValuePerc):z_{29}	1.5340	1.8995	0.3881	0.6718
s(zValuePerc):z_{49}	1.0000	1.0000	0.5952	0.4406
s(zValuePerc):z_{74}	2.1616	2.7315	0.7639	0.3898
fs(zValuePerc,sameValues,m=1,k=10)	170.1493	288.0000	2.2288	< 0.0001
fs(zValuePerc,trajectoryZ,m=1,k=7)	47.7523	1120.0000	0.0661	0.0001

Table 13: Coefficients of a generalized additive model with skew as the dependent variable with the value of z-variables at the point before [s] ceases from the output (WEAK). The model was fit with correction for autocorrelation with $\rho = 0.3$.

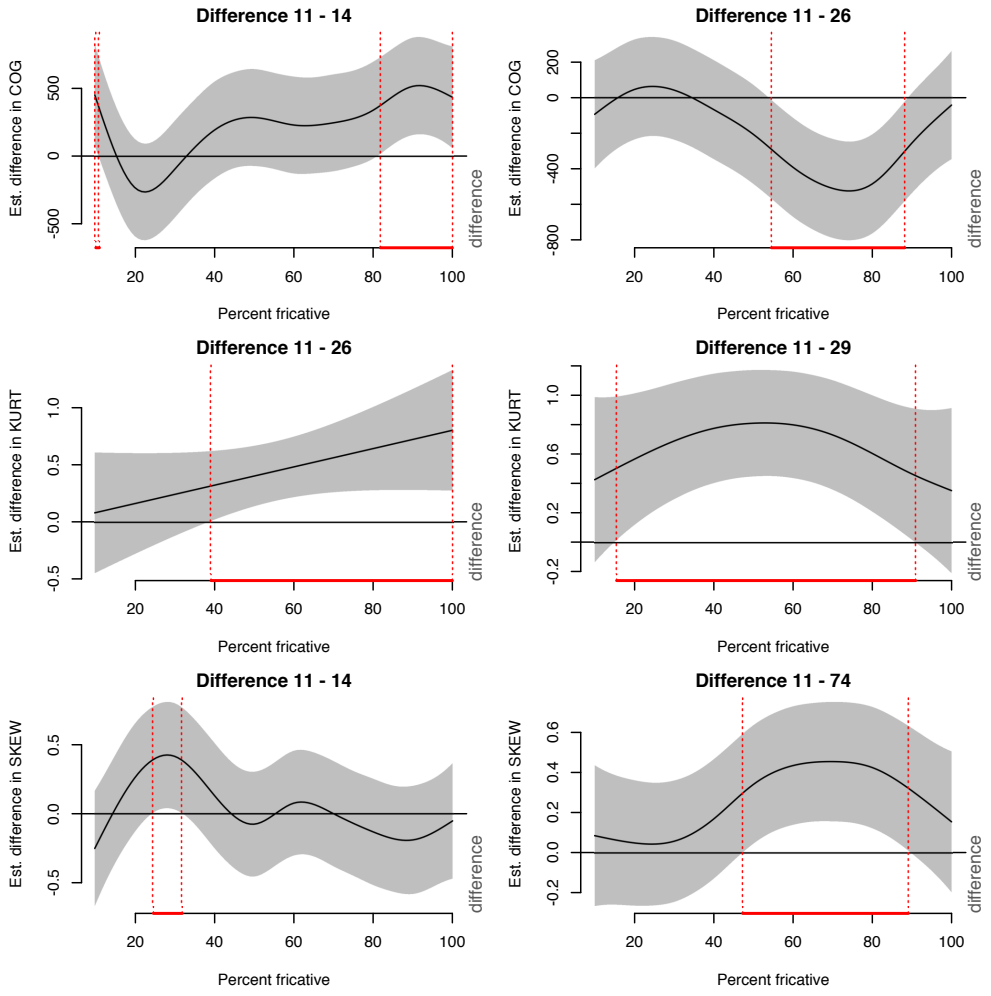

Figure 22: Pairwise difference smooths in COG, kurtosis, and skew between z_{11} and other two variables for models in Figure 21.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
URL http://tensorflow.org/

Abramson, A. S., Whalen, D., 2017. Voice onset time (vot) at 50: Theoretical and practical issues in measuring voicing distinctions. Journal of Phonetics 63, 75 – 86.
URL http://www.sciencedirect.com/science/article/pii/S0095447016301048

Alanazi, S., 2018. The Acquisition of English stops by Saudi L2 Learners. Ph.D. thesis, University of Essex.

Albright, A., Hayes, B., 2011. Learning and Learnability in Phonology. Wiley, Ch. 20, pp. 661–690.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444343069.ch20

Alderete, J., Tupper, P., 2018a. Connectionist approaches to generative phonology. In: Bosch, A., Hannahs, S. J. (Eds.), The Routledge Handbook of Phonological Theory. Routledge, New York, pp. 360–390.

Alderete, J., Tupper, P., 2018b. Phonological regularity, perceptual biases, and the role of phonotactics in speech error analysis. Wiley Interdisciplinary Reviews: Cognitive Science 9 (5), e1466.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1466

Alderete, J., Tupper, P., Frisch, S. A., 2013. Phonological constraint induction in a connectionist network: learning ocp-place constraints from data. Language Sciences 37, 52 – 69.
URL http://www.sciencedirect.com/science/article/pii/S0388000112001210

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning. pp. 214–223.

Avcu, E., 2018. Experimental investigation of the Subregular Hypothesis. In: Bennett, W. G., Hracs, L., Storoshenko, D. R. (Eds.), Proceedings of the 35th West Coast Conference on Formal Linguistics. Cascadilla, Somerville, MA, pp. 77–86.

Avcu, E., Shibata, C., Heinz, J., 2017. Subregular complexity and deep learning. In: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML).

Baayen, R. H., van Rij, J., de Cat, C., Wood, S. N., Jan 2016. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. arXiv e-prints, arXiv:1601.02043.

Barlow, J. A., 2001. Case study. Language, Speech, and Hearing Services in Schools 32 (4), 242–256.
URL https://pubs.asha.org/doi/abs/10.1044/0161-1461%282001/022%29

Bashivan, P., Kar, K., DiCarlo, J. J., 2019. Neural population control via deep image synthesis. Science 364 (6439).
URL https://science.sciencemag.org/content/364/6439/eaav9436

Beguš, G., 2018a. Post-nasal devoicing and the blurring process. Journal of Linguistics, 1–65.

Beguš, G., 2018b. Unnatural phonology: A synchrony-diachrony interface approach. Ph.D. thesis, Harvard University.

Boersma, P., Hayes, B., 2001. Empirical tests of the gradual learning algorithm. Linguistic Inquiry 32 (1), 45–86.
URL https://doi.org/10.1162/002438901554586

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., 1981. A note concerning /s/ plus stop clusters in the speech of language-delayed children. Applied Psycholinguistics 2 (1), 55–63.

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

Buchwald, A., Miozzo, M., 2012. Phonological and motor errors in individuals with acquired sound production impairment. Journal of Speech, Language, and Hearing Research 55 (5), S1573–S1586.
URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282012/11-0200%29

Bullinaria, J., 1997. Analyzing the internal representations of trained neural networks. In: Browne, A. (Ed.), Neural Network Analysis, Architectures and Algorithms. Institute of Physics Press, Bristol, pp. 3–26.

Catts, H. W., Jensen, P. J., 1983. Speech timing of phonologically disordered children. Journal of Speech, Language, and Hearing Research 26 (4), 501–510.
URL https://pubs.asha.org/doi/abs/10.1044/jshr.2604.501

Catts, H. W., Kamhi, A. G., 1984. Simplification of /s/ + stop consonant clusters. Journal of Speech, Language, and Hearing Research 27 (4), 556–561.
URL https://pubs.asha.org/doi/abs/10.1044/jshr.2704.556

Chandlee, J., 2014. Strictly local phonological processes. Ph.D. thesis, University of Delaware.

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

Clements, G. N., 1985. The geometry of phonological features. Phonology Yearbook 2 (1), 225–252.

Cohn, A. C., 2006. Is there gradient phonology? In: Fanselow, G., Féry, C., Schlesewsky, M. (Eds.), Gradience in Grammar: Generative Perspectives. Oxford University Press, Oxford, pp. 25–44.

Davis, S., Cho, M.-H., 2006. The distribution of aspirated stops and /h/ in American English and Korean: an alignment approach with typological implications. Linguistic 41 (4), 607–652.

de Lacy, P., 2006. Transmissibility and the role of the phonological component: A theoretical synopsis of Evolutionary Phonology. Theoretical Linguistics 32 (2), 185–196.

de Lacy, P., Kingston, J., 2013. Synchronic explanation. Natural Language and Linguistic Theory 31 (2), 287–355.

Dell, G. S., Juliano, C., Govindjee, A., 1993. Structure and content in language production: A theory of frame constraints in phonological speech errors. Cognitive Science 17 (2), 149 – 195.
URL http://www.sciencedirect.com/science/article/pii/0364021393900106

Donahue, C., McAuley, J., Puckette, M., 2019. Adversarial audio synthesis. In: ICLR.

Dresher, B. E., 2015. The motivation for contrastive feature hierarchies in phonology. Linguistic Variation 15 (1), 1–40.
URL https://www.jbe-platform.com/content/journals/10.1075/lv.15.1.01dre

Dupoux, E., 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. Cognition 173, 43 – 59.
URL http://www.sciencedirect.com/science/article/pii/S0010027717303013

Ernestus, M., 2011. Gradience and Categoricality in Phonological Theory. American Cancer Society, Ch. 89, pp. 1–22.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444335262.wbctp0089

Faruqui, M., Tsvetkov, Y., Neubig, G., Dyer, C., Jun. 2016. Morphological inflection generation using character sequence to sequence learning. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pp. 634–643.
URL https://www.aclweb.org/anthology/N16-1077

Fruehwald, J., 2016. The early influence of phonology on a phonetic change. Language 92 (2), 376–410.

Fruehwald, J., 2017. The role of phonology in phonetic change. Annual Review of Linguistics 3 (1), 25–42.
URL https://doi.org/10.1146/annurev-linguistics-011516-034101

Gerlach, S. R., 2010. The acquisition of consonant feature sequences: Harmony, metathesis and deletion patterns in phonological development. Ph.D. thesis, University of Minnesota.

Gibson, K. R., Tallerman, M., MacNeilage, P. F., 09 2012. The evolution of phonology. In: The Oxford Handbook of Language Evolution. Oxford University Press.
URL https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199541119.001.0001/oxfordhb-9780199541119-e-46

Glewwe, E., 2017. Substantive bias in phonotactic learning: Positional extension of an obstruent voicing contrast, talk presented at the 53rd meeting of *Chicago Linguistic Society*, Chicago, IL, May 25-27, 2017.

Glewwe, E., Zymet, J., Adams, J., Jacobson, R., Yates, A., Zeng, A., Daland, R., 2017. Substantive bias and word-final voiced obstruents: An artificial grammar learning study, talk presented at the 92nd Annual Meeting of the Linguistic Society of America, Salt Lake City, UT, January 4-7, 2018.

Goldwater, S., Johnson, M., 2003. Learning OT constraint rankings using a maximum entropy model. In: Spenader, J., Eriksson, A., Dahl, O. (Eds.), Proceedings of the Workshop on Variation within Optimality Theory. Stockholm University, Stockholm, pp. 111–20.

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

Guenther, F. H., 2016. Neural control of speech. MIT Press.

Guenther, F. H., Vladusich, T., 2012. A neural theory of speech acquisition and production. Journal of Neurolinguistics 25 (5), 408 – 422, is a neural theory of language possible? Issues from an interdisciplinary perspective.
URL http://www.sciencedirect.com/science/article/pii/S0911604409000682

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

Hale, M., Reiss, C., 2008. The phonological enterprise. Oxford University Press.

Haraguchi, Y., 2003. The acquisition of aspiration of voiceless stops and intonation patterns of english learners: Pilot study. In: Proceeding of the 8th conference of Pan-Pacific Association of Applied Linguistics. pp. 83–91.

Hayes, B., 1999. Phonetically-driven phonology: The role of Optimality Theory and inductive grounding. In: Darnell, M., Moravscik, E. (Eds.), Functionalism and Formalism in Linguistics, Volume I: General Papers. John Benjamins, Amsterdam, pp. 243–285.

Hayes, B., 2009. Introductory Phonology. Wiley-Blackwell, Malden, MA.

Hayes, B., White, J., 2013. Phonological naturalness and phonotactic learning. Linguistic Inquiry 44 (1), 45–75.

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

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

Heinz, J., 2011. Computational phonology – part ii: Grammars, learning, and the future. Language and Linguistics Compass 5 (4), 153–168.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2011.00268.x

Iverson, G. K., Salmons, J. C., 1995. Aspiration and laryngeal representation in germanic. Phonology 12 (3), 369–396.
URL http://www.jstor.org/stable/4420084

Jarosz, G., 2019. Computational modeling of phonological learning. Annual Review of Linguistics 5 (1), 67–90.
URL https://doi.org/10.1146/annurev-linguistics-011718-011832

Kamper, H., Elsner, M., Jansen, A., Goldwater, S., 2015. Unsupervised neural network based feature extraction using weak top-down constraints. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5818–5822.

Kawamoto, A. H., Liu, Q., Kello, C. T., 2015. The segment as the minimal planning unit in speech production and reading aloud: evidence and implications. Frontiers in Psychology 6, 1457.
URL https://www.frontiersin.org/article/10.3389/fpsyg.2015.01457

Keyser, S. J., Stevens, K. N., 2006. Enhancement and overlap in the speech chain. Language 82 (1), 33–63.

Kim, C.-W., 1970. A theory of aspiration. Phonetica 21, 107–116.

Kingston, J., Diehl, R. L., 1994. Phonetic knowledge. Language 70 (3), 419–454.

Kirby, J., Sonderegger, M., Jul. 2015. Bias and population structure in the actuation of sound change. arXiv e-prints, arXiv:1507.04420.

Kuhl, P. K., 2019/06/27 2010. Brain mechanisms in early language acquisition. Neuron 67 (5), 713–727.
URL https://doi.org/10.1016/j.neuron.2010.08.038

Lee, C.-y., Glass, J., Jul. 2012. A nonparametric Bayesian approach to acoustic model discovery. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Jeju Island, Korea, pp. 40–49.
URL https://www.aclweb.org/anthology/P12-1005

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

Legendre, G., Sorace, A., Smolensky, P., 2006. The Optimality Theory—Harmonic Grammar connection. In: Smolensky, P., Legendre, G. (Eds.), The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar. MIT Press, Cambridge, MA, pp. 339–402.

Lennes, M., 2003. f0-f1-f2-intensity_praat_script. praat script. Modified by Dan McCloy, Esther Le Grésauze, and Gašper Beguš.
URL https://depts.washington.edu/phonlab/resources/f0-F1-F2-intensity_praat_script.praat

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.
URL https://CRAN.R-project.org/doc/Rnews/

Lillicrap, T. P., Kording, K. P., Jul 2019. What does it mean to understand a neural network? arXiv e-prints, arXiv:1907.06374.

Lisker, L., 1984. How is the aspiration of english /p, t, k/ "predictable"? Language and Speech 27 (4), 391–394.
URL https://doi.org/10.1177/002383098402700409

Lowenstein, J. H., Nittrouer, S., 2008. Patterns of acquisition of native voice onset time in english-learning children. The Journal of the Acoustical Society of America 124 (2), 1180–1191.
URL https://doi.org/10.1121/1.2945118

Macken, M. A., Barton, D., 1980. The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. Journal of Child Language 7 (1), 41–74.

Macken, M. A., Ferguson, C. A., 1981. Phonological universals in language acquisition*. Annals of the New York Academy of Sciences 379 (1), 110–129.
URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1981.tb42002.x

Mahalunkar, A., Kelleher, J. D., 2018. Using regular languages to explore the representational capacity of recurrent neural architectures. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2018. Springer International Publishing, Cham, pp. 189–198.

Martin, A., Peperkamp, S., Dupoux, E., 2013. Learning phonemes with a proto-lexicon. Cognitive Science 37 (1), 103–124.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2012.01267.x

McLeod, S., van Doorn, J., Reed, V., 1996. Homonyms and cluster reduction in the normal development of children's speech. In: Proceedings of the Sixth Australian International Conference on Speech Science & Technology. pp. 331–336.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E. F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343 (6174), 1006–1010.
URL https://science.sciencemag.org/content/343/6174/1006

Moreton, E., 2008. Analytic bias and phonological typology. Phonology 25 (1), 83–127.

Moreton, E., Pater, J., 2012a. Structure and substance in artificial-phonology learning. Part I, Structure. Language and Linguistics Compass 6 (11), 686–701.

Moreton, E., Pater, J., 2012b. Structure and substance in artificial-phonology learning. Part II, Substance. Language and Linguistics Compass 6 (11), 702–718.

Moreton, E., Pater, J., Pertsova, K., 2017. Phonological concept learning. Cognitive Science 41 (1), 4–69.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12319

Nguyen, N., Delvaux, V., 2015. Role of imitation in the emergence of phonological systems. Journal of Phonetics 53, 46 – 54, on the cognitive nature of speech sound systems.
URL http://www.sciencedirect.com/science/article/pii/S0095447015000698

Odden, D., 2013. Formal phonology. Nordlyd 40 (1), 249–273.

Ohala, D. K., 1999. The influence of sonority on children's cluster reductions. Journal of Communication Disorders 32 (6), 397 – 422.
URL http://www.sciencedirect.com/science/article/pii/S0021992499000180

Oudeyer, P.-Y., 2001. Coupled neural maps for the origins of vowel systems. In: Proceedings of the International conference on artificial neural networks. Lecture notes in computer science. Springer, pp. 1171–1176, volume: 2130.

Oudeyer, P.-Y., 2002. Phonemic coding might result from sensory-motor coupling dynamics. MIT Press, pp. 406–416.
URL http://cogprints.org/2658/

Oudeyer, P.-Y., 2005. The self-organization of speech sounds. Journal of Theoretical Biology 233 (3), 435 – 449.
URL http://www.sciencedirect.com/science/article/pii/S0022519304005053

Oudeyer, P.-Y., 2006. Self-organization in the evolution of speech. Studies in the evolution of language ; 6. Oxford University Press, Oxford.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., Chang, E. F., 01 2012. Reconstructing speech from human auditory cortex. PLOS Biology 10 (1), 1–13.
URL https://doi.org/10.1371/journal.pbio.1001251

Pater, J., 2009. Weighted constraints in generative linguistics. Cognitive Science 33, 999–1035.

Pater, J., 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. Language.

Pierrehumbert, J., 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In: Bybee, J. L., Hopper, P. J. (Eds.), Frequency Effects and the Emergence of Lexical Structure. John Benjamins, Amsterdam, pp. 137–157.

Pouplier, M., Marin, S., Waltl, S., 2014. Voice onset time in consonant cluster errors: Can phonetic accommodation differentiate cognitive from motor errors? Journal of Speech, Language, and Hearing Research 57 (5), 1577–1588.
URL https://pubs.asha.org/doi/abs/10.1044/2014_JSLHR-S-12-0412

Prickett, B., Traylor, A., Pater, J., 2019. Learning reduplication with a variable-free neural network, ms., University of Massachusetts, Amherst. http://works.bepress.com/joe_pater/38/ (accessed 23 May 2019).

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

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

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

Rawski, J., Heinz, J., 2019. No free lunch in linguistics or machine learning: Response to pater. Language.

Rentz, B., 2017. spectral_moments.praat. praat script.
URL https://github.com/rentzb/praat-scripts/blob/master/spectral_moments.praat

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

Saffran, J. R., Aslin, R. N., Newport, E. L., 1996. Statistical learning by 8-month-old infants. Science 274 (5294), 1926–1928.
URL https://science.sciencemag.org/content/274/5294/1926

Saffran, J. R., Werker, J. F., Werner, L. A., 2007. The Infant's Auditory World: Hearing, Speech, and the Beginnings of Language. American Cancer Society, Ch. 2.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470147658.chpsy0202

Schatz, T., Feldman, N., Goldwater, S., Cao, X. N., Dupoux, E., May 2019. Early phonetic learning without phonetic categories – insights from machine learning.
URL psyarxiv.com/fc4wh

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

Silfverberg, M. P., Mao, L., Hulden, M., 2018. Sound analogies with phoneme embeddings. In: Proceedings of the Society for Computation in Linguistics (SCiL) 2018. pp. 136–144.
URL https://www.aclweb.org/anthology/W18-0314

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

Smolensky, P., Legendre, G., 2006. The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1 & 2. MIT press.

Sóskuthy, M., Mar 2017. Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. arXiv e-prints, arXiv:1703.05339.

Syrika, A., Nicolaidis, K., Edwards, J., Beckman, M. E., 2011. Acquisition of initial /s/-stop and stop-/s/ sequences in Greek. Language and Speech 54 (3), 361–386, pMID: 22070044.
URL https://doi.org/10.1177/0023830911402597

Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., Dupoux, E., 2015. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In: Proceedings of Interspeech.

Trubetzkoy, N. S., 1939. Grundzüge der Phonologie. Travaux de Cercle linguistique de Prague ; 7. [s.n.] ;, Prague.

Vaux, B., 2002. Aspiration in English, ms., Harvard University. Accessed on June 27, 2019.
URL https://www.academia.edu/300605/Aspiration_In_English

Vaux, B., Samuels, B., 2005. Laryngeal markedness and aspiration. Phonology 22 (3), 395–436.

Weber, N., Shekhar, L., Balasubramanian, N., 2018. The fine line between linguistic generalization and failure in Seq2Seq-attention models. In: Proceedings of the Workshop on Generalization in the Age of Deep Learning. Association for Computational Linguistics, New Orleans, Louisiana, pp. 24–27.
URL https://www.aclweb.org/anthology/W18-1004

White, J., 2014. Evidence for a learning bias against saltatory phonological alternations. Cognition 130 (1), 96 – 115.
URL http://www.sciencedirect.com/science/article/pii/S0010027713001923

White, J., 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. Language 93 (1), 1–36.

Wilson, C., 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. Cognitive Science 30, 945–982.

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

Yildiz, Y., 2005. The structure of initial/s/-clusters: evidence from L1 and L2 acquisition. In: Tzakosta, M., Levelt, C., van der Weijer, J. (Eds.), Developmental Paths in Phonological Acquisition. pp. 163–187, special issue of Leiden Papers in Linguistics 2.1.

Young, E. D., 2008. Neural representation of spectral and temporal information in speech. Philosophical Transactions of the Royal Society B: Biological Sciences 363 (1493), 923–945.
URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2007.2151

Zuidema, W., de Boer, B., 2009. The evolution of combinatorial phonology. Journal of Phonetics 37 (2), 125 – 144.
URL http://www.sciencedirect.com/science/article/pii/S0095447008000624

Version 3: August 1, 2019
Version 2: July 6, 2019
Version 1: May 28, 2019