frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 08 July 2020
doi: 10.3389/frai.2020.00044

Generative Adversarial Phonology: Modeling Unsupervised Phonetic and Phonological Learning With Neural Networks

Gašper Beguš [1,2]*

[1] Department of Linguistics, University of California, Berkeley, Berkeley, CA, United States, [2] Department of Linguistics, University of Washington, Seattle, WA, United States

Training deep neural networks on well-understood dependencies in speech data can provide new insights into how they learn internal representations. This paper argues that acquisition of speech can be modeled as a dependency between random space and generated speech data in the Generative Adversarial Network architecture and proposes a methodology to uncover the network's internal representations that correspond to phonetic and phonological properties. 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. A Generative Adversarial Network was trained on an allophonic distribution in English, in which voiceless stops surface as aspirated word-initially before stressed vowels, except if preceded by a sibilant [s]. The network successfully learns the allophonic alternation: the network's generated speech signal contains the conditional distribution of aspiration duration. The paper proposes a technique for establishing the network's internal representations that identifies latent variables that correspond to, for example, presence of [s] and its spectral properties. By manipulating these variables, we actively control the presence of [s] and its frication amplitude in the generated outputs. This suggests that the network learns to use latent variables as an approximation of phonetic and phonological representations. Crucially, we observe that the dependencies learned in training extend beyond the training interval, which allows for additional exploration of learning representations. The paper also discusses how the network's architecture and innovative outputs resemble and differ from linguistic behavior in language acquisition, speech disorders, and speech errors, and how well-understood dependencies in speech data can help us interpret how neural networks learn their representations.

Keywords: generative adversarial networks, deep neural network interpretability, language acquisition, speech, voice onset time, allophonic distribution

OPEN ACCESS

Edited by:
Kemal Oflazer, Carnegie Mellon University in Qatar, Qatar

Reviewed by:
David R. Mortensen, Carnegie Mellon University, United States
Robert Malouf, San Diego State University, United States
Kevin Tang, University of Florida, United States

***Correspondence:**
Gašper Beguš
begus@berkeley.edu

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *28 January 2020*
Accepted: *19 May 2020*
Published: *08 July 2020*

Citation:
Beguš G (2020) Generative Adversarial Phonology: Modeling Unsupervised Phonetic and Phonological Learning With Neural Networks. Front. Artif. Intell. 3:44. doi: 10.3389/frai.2020.00044

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 a grammar that enables them to do

so (Saffran et al., 1996, 2007; Kuhl, 2010). In addition to syntactic, morphological, and semantic representations, the learner needs to learn phonetic representations and phonological grammar: to analyze and in turn produce speech as a continuous acoustic stream represented by 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 a 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; Moreton, 2008; Moreton and Pater, 2012a,b; de Lacy and Kingston, 2013; 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). 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 already operates on the abstract, discrete phonological level. The models thus require strong assumptions that phonetic learning has already taken place, and that phonemes as discrete units have already been inferred from continuous speech data (for an overview of the literature, see Oudeyer, 2005, 2006; Dupoux, 2018).

This paper proposes that language acquisition can be modeled with Generative Adversarial Networks (Goodfellow et al., 2014). More specifically, phonetic and phonological computation is modeled as the mapping from random space to generated data of a Generative Adversarial Network (Goodfellow et al., 2014) trained on raw unannotated acoustic speech data in an unsupervised manner (Donahue et al., 2019). To the author's knowledge, language acquisition has not been modeled within the GAN framework despite several advantages of this architecture. The characteristic feature of the GAN architecture is an interaction between the Generator network that outputs raw data and the Discriminator that distinguishes real data from Generator's outputs (Goodfellow et al., 2014). A major advantage of the GAN architecture is that learning is completely unsupervised, the networks include no language-specific elements, and that, as is argued in Section 4 below, phonetic learning is modeled simultaneously 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 phonological learning is likely one step closer to reality than models that operate with symbolic computation and assume phonetic learning has already taken place independently of phonology (and vice versa).

We argue that the latent variables in the input of the Generator network can be modeled as approximates to phonetic or potentially phonological representations 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 thus has a parallel in human speech: the imitation principle (Nguyen and Delvaux, 2015). The Discriminator's function is to enforce that the Generator's outputs resemble (but do not replicate) the inputs as closely as possible. The GAN network thus incorporates both the pre-articulatory production elements (the Generator) as well as the imitation principle (the Discriminator) in speech acquisition. While other neural network architectures might be appropriate for modeling phonetic and phonological learning as well, the GAN architecture is unique in that it combines a network that produces innovative data (the Generator) with a network that forces imitation in the Generator. Unlike, for example, autoencoder networks, the Generative Adversarial network lacks a direct connection between the input and output data and generates innovative data rather than data that resembles the input as closely as possible.

We train a Generative Adversarial Network architecture implemented for audio files in Donahue et al. (2019) (WaveGAN) on 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 for non-desired effects, which is why only sequences of the shape #TV and #sTV[1] are fed to the model. This allophonic distribution is appropriate for testing learnability in a GAN architecture, 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. To our knowledge, this is the first proposal that tests whether neural networks are able to learn an allophonic distribution based on raw acoustic data.

The hypothesis of the computational experiment presented in Section 4 is the following: if VOT duration is conditioned on the presence of [s] 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. Because the allophonic distribution is not strictly local and has to be learned and actively controlled by speakers (i.e., is not automatic), evidence for this type of learning is considered phonological learning in the broadest sense. Conditioning the presence of a phonetic feature based on the presence or absence of a phoneme that is not automatic is, in most models, considered part of phonology and is derived with

[1]T represents voiceless stops /p, t, k/, V represents vowels (see **Figure 4**), and # represents a word boundary.

phonological computation. That the tested distribution is non-automatic and has to be actively controlled by the speakers is evident from L1 acquisition: failure to learn the distribution results in longer VOT durations in the sT condition documented in L1 acquisition (see Section 5.1).

The results suggest that phonetic and phonological learning can be modeled simultaneously, without supervision, 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 outputs that contain this allophonic distribution (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 the GAN's generated output. As argued in Section 4.2, some outputs are consistent with human linguistic behavior and suggest that the model recombines individual sounds, resembling phonemic representations and productivity in human language acquisition (Section 5).

This paper also proposes a technique for establishing the Generator's internal representations. The inability to uncover networks' representations has been used as an argument against neural network approaches to linguistic data (among others in Rawski and Heinz, 2019). We argue that the internal representation of a network can be, at least partially, uncovered. By regressing annotated dependencies between the Generator's latent space and output data, we identify values in the latent space that correspond to linguistically meaningful features in generated outputs. This paper demonstrates that manipulating the chosen values in the latent space has phonetic effects in the generated outputs, such as the presence of [s] and the amplitude of its frication. In other words, the GAN learns to use random noise as an approximation of phonetic (and potentially phonological) representations. This paper proposes that dependencies, learned during training in a latent space that is limited by some interval, extend beyond this interval. This crucial step allows for the discovery of several phonetic properties that the model learns.

By modeling phonetic and phonological learning with neural networks without any language-specific assumptions, the paper also addresses a broader question of how many language-specific elements are needed 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 approaches or pre-determined constraints with features and feature matrices in connectionist approaches. The model proposed here lacks language-specific assumptions (similar to the exemplar-based models). Comparing the performance of substance-free models with competing proposals and human behavior should result in a better understanding of what aspects of phonological grammar and acquisition are domain-specific (Section 5).

In the following, we first survey existing theories of phonological grammar and literature on computational approaches to phonology (Section 2). In Section 3, we present the model in Donahue et al. (2019) based on Radford et al. (2015) and provide acoustic and statistical analysis of the training data. The network's outputs are first acoustically analyzed and described in Sections 4.1 and 4.2. In Section 4.3, we present a technique for establishing the network's internal representations and test it with two generative tests. In Section 4.5, we analyze phonetic properties of the network's internal representations. Section 5 compares the outputs of the model with L1 acquisition, speech impairments, and speech errors.

2. PREVIOUS WORK

In the generative tradition, *phonological grammar* derives surface phonetic outputs from phonological inputs (Chomsky and Halle, 1968). For example, /p/ is an abstract unit that can surface (be realized) with variations at 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 English phoneme /p/ as an abstract mental unit is unspecified for aspiration (or absence thereof) in the underlying representation (/ˈpɪt/ "pit" and /ˈspɪt/ "spit"). The surface phonetic outputs after the phonological derivation had taken place are [ˈp^hɪt] with the aspiration and [ˈspɪt] without the aspiration.

One of the main objectives 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 and Optimality Theory. The first approach uses rewrite rules (Chomsky and Halle, 1968) or finite state automata (Heinz, 2010; Chandlee, 2014) to derive outputs from inputs through derivation. A connectionist approach called Optimality Theory (Prince and Smolensky, 2004) and related proposals such as Harmonic Grammar and Maximum Entropy (MaxEnt) grammar (Legendre et al., 1990, 2006; Goldwater and Johnson, 2003; Wilson, 2006; Hayes and Wilson, 2008; Pater, 2009; Hayes and White, 2013; White, 2014, 2017), on the other hand, model phonological grammar as input-output pairing: the grammar chooses the most optimal output given an input. These models were heavily influenced by the early advances in neural network research (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; Kirov, 2017; Alderete and Tupper, 2018a; Dupoux, 2018; Mahalunkar and Kelleher, 2018; Weber et al., 2018; Prickett et al., 2019, for cautionary notes, see Rawski and Heinz, 2019). One of the promising implications of neural network modeling is the ability to test generalizations that the models produce without language-specific assumptions (Pater, 2019).

In opposition to the generative approaches, there exists a long tradition of usage-based models in phonology (Bybee, 1999; Silverman, 2017) which diverges from the generative approaches in some crucial aspects. Exemplar models (Johnson, 1997, 2007; Pierrehumbert, 2001; Gahl and Yu, 2006; Kaplan, 2017), for example, assume that phonetic representations are stored as experiences or exemplars. Grammatical behavior emerges as a consequence of generalization (or computation) over a cloud of exemplars (Johnson, 2007; Kaplan, 2017).

In this framework, there is no direct need for a separate underlying representation from which the surface outputs are derived (or optimized). Several phenomena in phonetics and phonology have been successfully derived within this approach (for an overview, see Kaplan, 2017), and the framework allows phonology to be modeled computationally. The computational models often involve interacting agents learning some simplified phonetic properties (e.g., de Boer, 2000; Wedel, 2006; Kirby and Sonderegger, 2015).

The majority of existing computational models in phonology (including finite state automata, the MaxEnt model and the existing neural network methods) 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 that can be approximated as phonemes. This means that either 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; Futrell et al., 2017; Kirov, 2017; Prickett et al., 2019) and featural organizations (Faruqui et al., 2016; Silfverberg et al., 2018). The existing models also require strong assumptions about learning: underlying representations, for example, are pre-assumed and not inferred from data (Kirov, 2017; Prickett et al., 2019).

Most models in the subset of the proposals that operate with continuous phonetic data 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—i.e., allophonic distributions. Phonological learning in most of these proposals is thus modeled as if phonetic learning (or at least a subset of phonetic learning) has already taken place: the initial state already includes phonemic inventories, phonemes as discrete units, feature matrices that have 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 (Lee and Glass, 2012; Schatz et al., 2019). 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 proposal 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.

A number of earlier works in the connectionist approach included basic neural network architectures to model mapping from some simplified phonetic space to the discrete phonological space (McClelland and Elman, 1986; Gaskell et al., 1995; Plaut and Kello, 1999; Kello and Plaut, 2003). Input to most of these models is not raw acoustic data (except in Kello and Plaut, 2003), but already extracted features. Learning in these models is also not unsupervised: the models come pre-specified with discretized phonetic or phonological units. None of the models are generative and do not model learning of phonological processes, but rather of classifying a simplified phonetic space with already available phonological elements.

Recently, neural network models for unsupervised feature extraction have seen success in modeling acquisition of phonetic features from raw acoustic data (Räsänen et al., 2016; Eloff et al., 2019; Shain and Elsner, 2019). 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 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, do not show a direct relationship between individual variables in the latent space and acoustic outputs (as in Section 4.4 and **Figure 14**), 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 argued below, the model based on a Generative Adversarial Network (GAN) learns not only to generate innovative data that closely resemble human speech, but also learns internal representations that resemble phonological learning with unsupervised phonetic learning from raw acoustic data. Additionally, the model is generative and outputs both the conditional allophonic distributions and innovative data that can be compared to productive outputs in human speech acquisition.

3. MATERIALS

3.1. The Model: Donahue et al. (2019) Based on Radford et al. (2015)

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 to distinguish real data

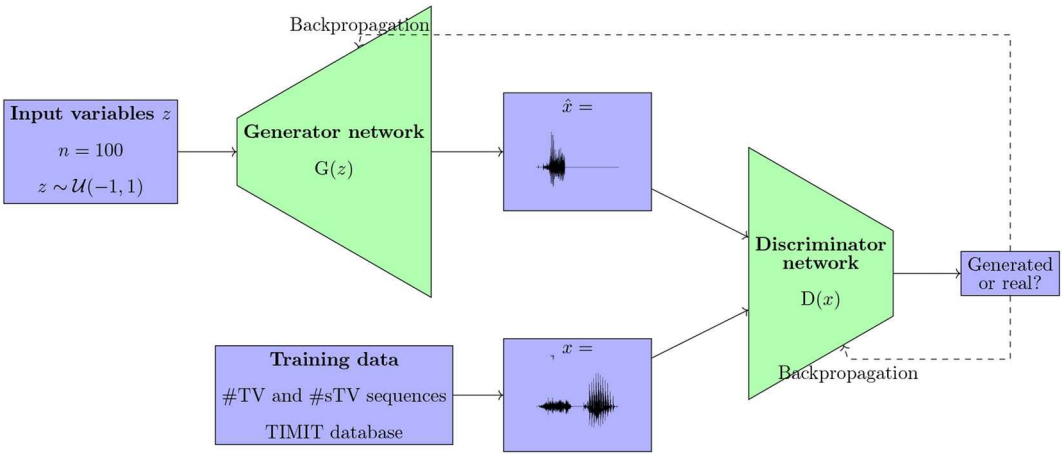

FIGURE 1 | A diagram showing the Generative Adversarial architecture as proposed in Goodfellow et al. (2014) and Donahue et al. (2019), trained on data from the TIMIT database in this paper.

from the outputs of the Generator network (**Figure 1**). The Generator is trained to generate data that maximizes the error rate 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 the real data (**Figure 1**).

Applying the GAN architecture to time-series data such as a continuous speech stream poses 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 along with the code in Donahue et al. (2019) were used for training in this paper. The model takes 1 s 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, the 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 (Donahue et al., 2019). 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 files) as its input and outputs a single value. The Discriminator's weights are updated five times per each update of the Generator. The initial GAN design as proposed by Goodfellow et al. (2014) trained the Discriminator network to distinguish real from generated data. Training such models, however, posed substantial challenges (Donahue et al., 2019). Donahue et al. (2019) implement the 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). For exact dimensions of each layer and other details of the model, see Donahue et al. (2019).

3.2. Training Data

The model was trained on the 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 as unaspirated [p, t, k] (Vaux and Samuels, 2005). A minimal pair illustrating this allophonic distribution is [ˈp^hɪt] "pit" vs. [ˈspɪt] "spit." The most prominent phonetic correlate of this allophonic distribution is the difference in Voice Onset Time (VOT) duration (Lisker and Abramson, 1964; Abramson and Whalen, 2017) between the aspirated and unaspirated voiceless stops. VOT is the duration between the release of the stop ([p, t, k]) and the onset of periodic vibration in the following vowel.

The model was trained on data from the TIMIT database (Garofolo et al., 1993)[2]. 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 6,300 sentences, 10 sentences per 630 speakers from 8 major dialectal areas in the US (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

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

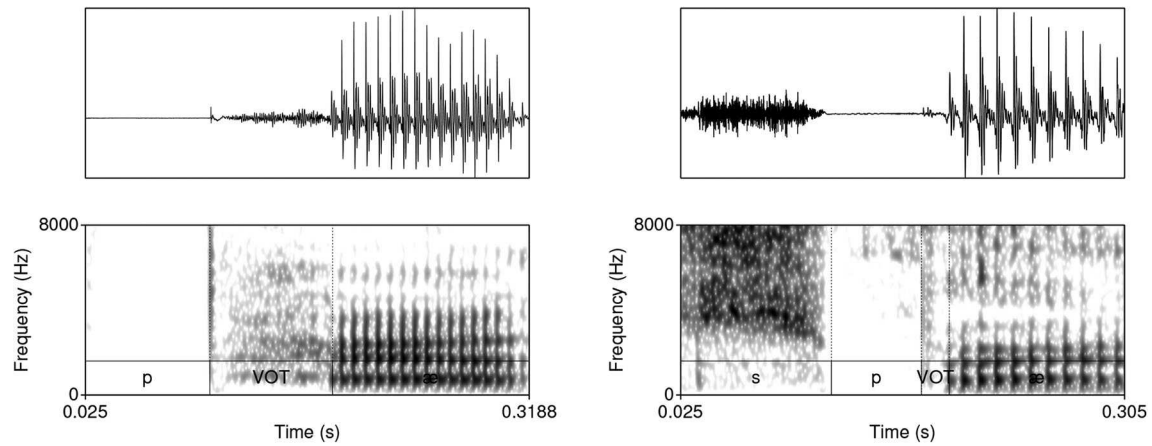

FIGURE 2 | Waveforms and spectrograms (0 – 8,000 Hz) of [$p^h æ$] **(left)** and [spæ] **(right)** illustrating typical training data with annotations from TIMIT. Only the raw audio data (in ·wav format) were used in training. The annotation illustrates a substantially longer duration of VOT in word-initial stops when no [s] precedes.

by a vowel (#sTV). The training data includes 4,930 sequences with the structure #TV (90.2%) and 533 (9.8%) 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: the absence/presence of [s] is interrupted from the VOT duration by a period of closure in the training data. That VOT is significantly shorter if T is preceded by [s] in the training data is confirmed by a Gamma regression model: $\beta = -0.84, t = -49.69, p < 0.0001$ (for details, see Section 1, **Supplementary Materials**).

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 model more challenging as well as more realistic. 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 an unstressed vowel). Violin plots in **Figure 3** illustrate that aspiration of stops before an unstressed vowel can be as short as in the #sTV condition. 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.

The training data is not completely naturalistic: #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 acoustic inputs. If the entire lexicon was 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: the presence or absence of [s]. It is thus less likely that the distribution of VOT duration across the main condition of interest, the 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 4**, 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 problematic for generated inputs with fewer training steps. Because the primary interest of the experiment is the difference in VOT durations between two groups (the 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.

4. EXPERIMENT

4.1. Training and Generation

The purpose of this paper is to model phonetic and phonological learning. For this reason, the data was generated and examined at different points as the Generator network was in the process of being trained. For the purpose of modeling learning, it is more informative to probe the networks with fewer training steps, which allows a comparison between the model's outputs and L1 acquisition (Section 5.1). Outputs of the network are analyzed after 12,255 steps (Section 4.2). The number of steps was chosen as a compromise between quality of output data and the number of epochs in the training. Establishing the number of training steps at which an effective acoustic analysis can be performed is at this point somewhat arbitrary. We generated outputs of the Generator model trained after 1,474, 4,075, 6,759, 9,367, and 12,255 steps and manually inspected them. The model trained after 12,255 steps was considered the first that allowed

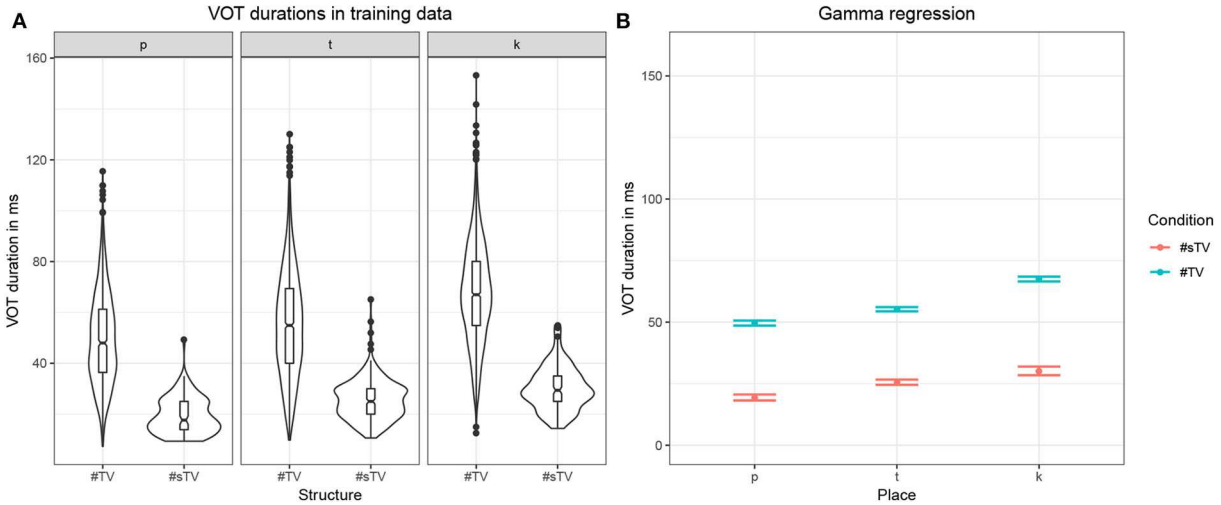

FIGURE 3 | (A) Violin plots with box-plots of durations in ms of VOT in the training data based on two conditions: when word-initial #TV sequence is not preceded by [s] (#TV) and when it is preceded by [s] (#sTV) across the three places of articulation: [p], [t], [k]. **(B)** Fitted values of VOT durations with 95% confidence intervals from the Gamma (with log-link) regression model in **Table S1** in **Supplementary Materials**.

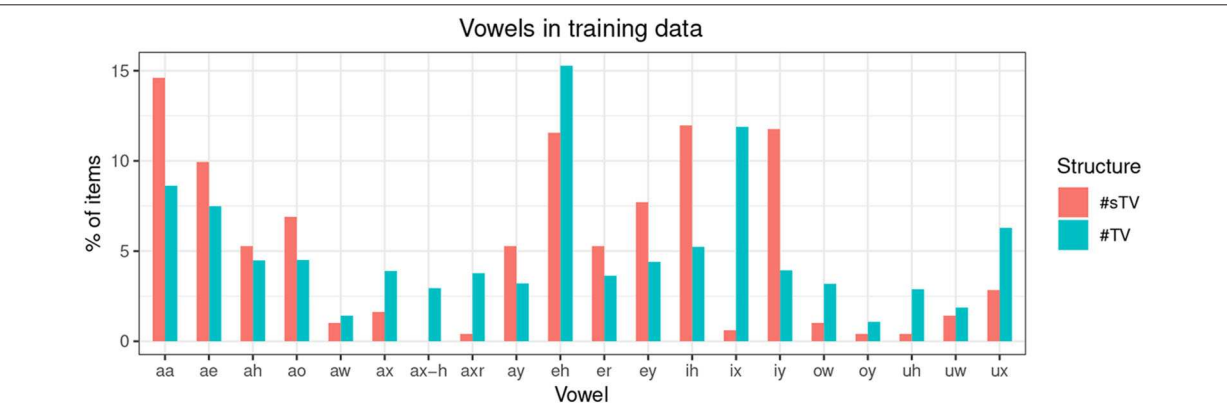

FIGURE 4 | Distribution of training items according to vowel identity as described in TIMIT in ARPABET, where aa = ɑ, ae = æ, ah = ʌ, ao = ɔ, aw = aʊ, ax = ə, ax-h = ə̥, axr = ɚ, ay = aɪ, eh = ɛ, er = ɝ, ey = eɪ, ih = ɪ, ix = ɨ, iy = i, ow = oʊ, oy = oɪ, uh = ʊ, uw = u, ux = ʉ in the International Phonetic Alphabet.

a reliable acoustic analysis based on quality of the generated outputs. It would be informative to test how accuracy of labeled data improves with training steps, but this is left for future work. The model was trained on a single NVIDIA K80 GPU. The network was trained at an approximate pace of 40 steps per 300 s. In Section 4.2, we present measurements of VOT durations in the #sTV and #TV conditions in the generated outputs and discuss linguistically interpretable innovative outputs that violate the training data. In Section 4.3.1, we propose a technique for recovering the Generator network's internal representations; in Section 4.4 we illustrate that manipulating these variables has a phonetically meaningful effect in the output data.

4.2. VOT Duration

The Generator network after 12,255 steps (∼ 716 epochs) generates acoustic data that appear close to actual speech data. **Figure 5** illustrates a typical 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, 2,000 samples were generated and manually inspected. First, VOT duration was manually annotated in all #sTV sequences. There were altogether 156 such sequences. To perform significance testing on a similar sample size, the first 158 sequences of the #TV structure were also annotated for VOT duration. VOT was measured from the release of closure to the onset of periodic vibration with clear formant structure. Altogether 314 generated samples were thus annotated. Only samples with structure that resembles real acoustic outputs and for which VOT could be determined were annotated. The proportion of inputs for which a clear #sTV or #TV sequence was not recognizable is relatively small: in only 8 of the first 175 annotated outputs (4.6%) was it not possible to estimate the VOT duration or whether the sequence is of the #TV or #sTV

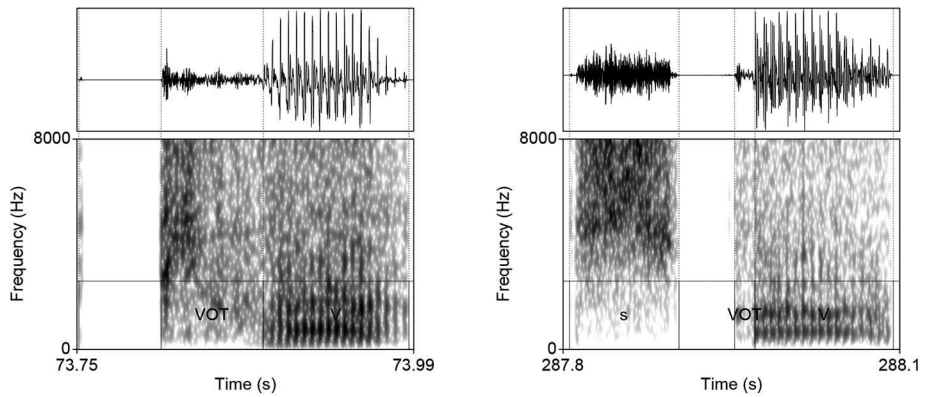

FIGURE 5 | Waveforms and spectrograms (0–8,000 Hz) of typical generated samples of #TV **(left)** and #sTV **(right)** sequences from a Generator trained after 12,255 steps.

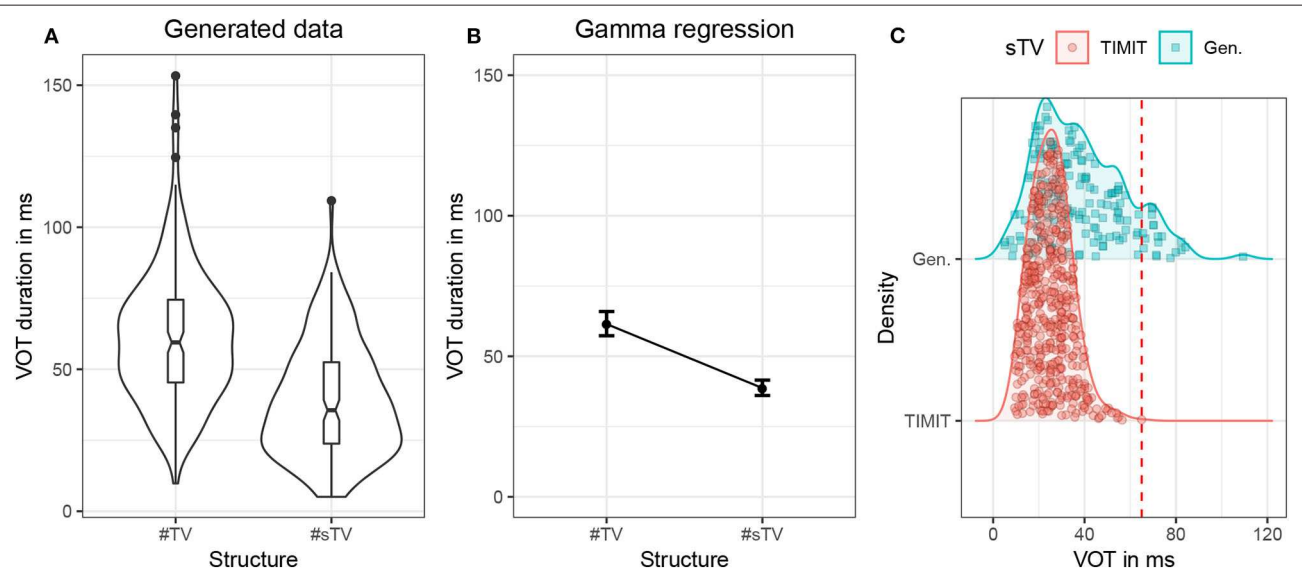

FIGURE 6 | (A) Violin plots with box-plots of durations in ms of VOT in the generated data based on two conditions: when word-initial #TV sequence is not preceded by [s] (#TV) and when it is preceded by [s] (#sTV). **(B)** Estimates of VOT duration based on fitted values with 95% confidence intervals across two conditions, #TV and #sTV in the generated data. **(C)** Duration (in ms) of VOT in the #sTV condition that compares the TIMIT training data (red circles) and generated outputs (green squares). The dashed line represents the longest VOT duration in the training data. The figure illustrates the proportion of outputs that violate the training data.

structure. **Figure 6** shows the raw distribution of VOT durations in the generated samples that closely resembles the distribution in the training data (**Figure 3**).

The results suggest that the network does learn the allophonic distribution: VOT duration is significantly shorter in the #sTV condition ($\beta = -2.79, t = -78.34, p < 0.0001$; for details of the statistical model, see Section 2, **Supplementary Materials**). **Figure 6** illustrates estimates of VOT duration across the two conditions with 95% confidence intervals. The model, however, shows clear traces that the learning is imperfect and that the generator network fails to learn the distribution *categorically*. This is strongly suggested by the fact that VOT durations are substantially longer in the generated data compared to the training data. The difference in means between the #TV and #sTV conditions in the training data is 32.35 ms, while in the generated data the difference is 22.52 ms. The ratio between the two conditions in the training data is 2.34, while the generated data's ratio is 1.59.

Another aspect of generated data that also strongly suggests the learning is imperfect is the fact that the longest VOT durations in the #sTV condition in the generated data are substantially longer than the longest VOT durations in the training data, where the longest duration reaches 65 ms (see **Table 1** and **Figure 3**). VOT in the generated data is in 19 out of 156 total #sTV sequences (or 12.2%) longer than 65.5 ms, the longest VOT in the training data. The longest three VOT durations in #sTV sequences are, for example, 109.35, 84.17, and 82.37 ms.

TABLE 1 | Raw VOT durations in ms for the training data with SD and Range.

Structure	Place	VOT	SD	Lowest	Highest	Count
#TV	p	49.6	18.0	7.3	115.5	1,018
	t	55.2	20.7	9.8	130.0	1,799
	k	67.5	19.5	12.5	153.1	2,112
#sTV	p	19.4	7.1	9.4	49.2	115
	t	25.6	7.9	10.6	65.0	288
	k	30.1	8.6	14.4	55.0	130

This generalization holds also in proportional terms. To control for the overall duration of segments in the output, we measure ratio of VOT duration and duration of the preceding [s] (i.e., thus controlling for "speech rate"). The highest ratio between the VOT duration and the duration of preceding [s] ($\frac{\text{VOT}}{\text{[s]}}$) in the training data is 0.77[3], which appears 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. Several values of the ratios between VOT and [s] duration in the generated data are substantially higher compared to the training data. In the three outputs with longest absolute duration of VOT, the ratios are 1.91, 1.40, and 0.89. Other high ratios measured include, for example, 1.79, 1.72, 1.60, 1.50, 1.46. **Figure 7** shows two such cases. It is clear that the generator fails to reproduce the conditioned durational distribution from the training data in these particular cases. In other words, while the Generator learns to output significantly shorter VOT durations when [s] is present in the output, it occasionally (in approximately 12.2% of cases) fails to observe this generalization and outputs a long VOT in the #sTV condition which is longer than any VOT duration in the #sTV condition in the training data. As will be argued in Section 5.1, the outcomes of this imperfect learning closely resemble L1 acquisition.

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. Among approximately 3,000 generated samples analyzed, we observe generated outputs that feature only frication noise of [s] and periodic vibration of the following vowel, but lack stop elements completely (e.g., closure and release of the stop). In other words, the generator occasionally outputs a linguistically valid and innovative #sV sequence for which no evidence was available in the training data. Such innovative sequences in which the segments are omitted or inserted are rare compared to innovative outputs with longer VOT—approximately two per 3,000 inspected cases (but the overall rate of outputs that are acoustically difficult to analyze is also small: 4.6%). All sequences containing [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. The minimal duration of closure in #sTV sequences in the training data is 9.2 ms, and the minimal duration of VOT is 9.4 ms. Aspiration noise in stops that resembles frication of [s] and homorganic sequences of [s] followed by an alveolar stop [t] (#stV) are occasionally acoustically similar to the sequence without the stop (#sV) due to similar articulatory positions or because frication noise from [s] carries onto the homorganic alveolar closure which can be very short. Such data points in the training data can serve as the basis for the innovative output #sV. However, there is a clear fall and a second rise of noise amplitude after the release of the stop in #stV sequences. **Figure 8** shows two cases of the Generator network outputting an innovative #sV sequence without any stop-like fall of the amplitude, for which no direct evidence exists in the training data.

Similarly, the Generator occasionally outputs a sequence with two stops and a vowel (#TTV). One potential source of such innovative sequences might be residual noise that is sometimes present during the period of closure in the TIMIT database. However, residual noise in the training data differs substantially from a clear aspiration noise in the generated #TTV sequences. **Figure 9** illustrates two generated 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[4]. 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 risk with overfitting in a GAN architecture is that the Generator network would learn to fully replicate the input[5]. The best evidence against overfitting is precisely the fact that the Generator network outputs samples that substantially violate data distributions (**Figures 7–9**)[6].

4.3. Establishing Internal Representations

Establishing what and how neural networks learn is a challenging task (Lillicrap and Kording, 2019). Exploration of latent space in the GAN architecture has been performed before (Radford et al., 2015; Donahue et al., 2017; Lipton and Tripathi, 2017), but to the author's knowledge, most previous work did not focus on discovering meaningful values (phonetic correlates) of each variable and has not fully used the potential to extend those variables to values well outside the training range (15 or 25). Below, we propose a technique for uncovering dependencies

[3]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 1.91 in the generated data is still substantially higher than the 1.17.

[4]For evidence that units smaller than segments are phonologically relevant, see Inkelas and Shih (2017) and literature therein.

[5]In general, GANs do not overfit (Adlam et al., 2019; Donahue et al., 2019), as is suggested by our data. Even if overfitting did occur, it would result from training a Generator without a direct access to the training data (unlike in the autoencoder models, where the input training data and outputs are directly connected).

[6]Donahue et al. (2019) test overfitting on models trained with a substantially 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.

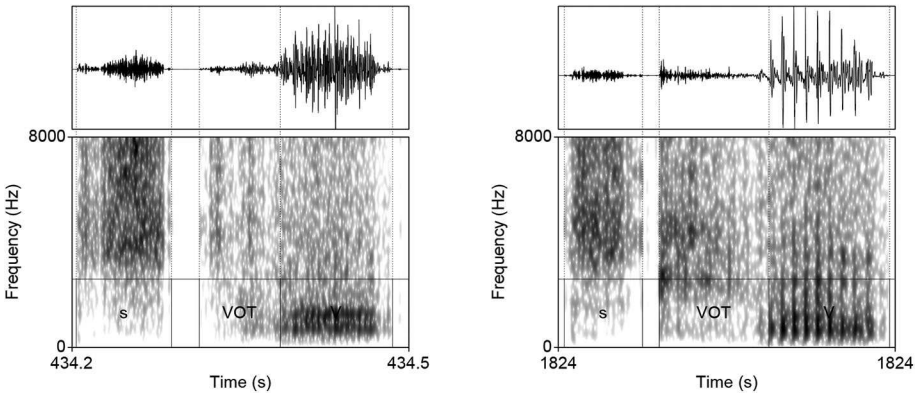

FIGURE 7 | Waveforms and spectrograms (0–8,000 Hz) of two generated outputs of #sTV sequences in which the stop has longer VOT than any VOT in #sTV condition in the training data.

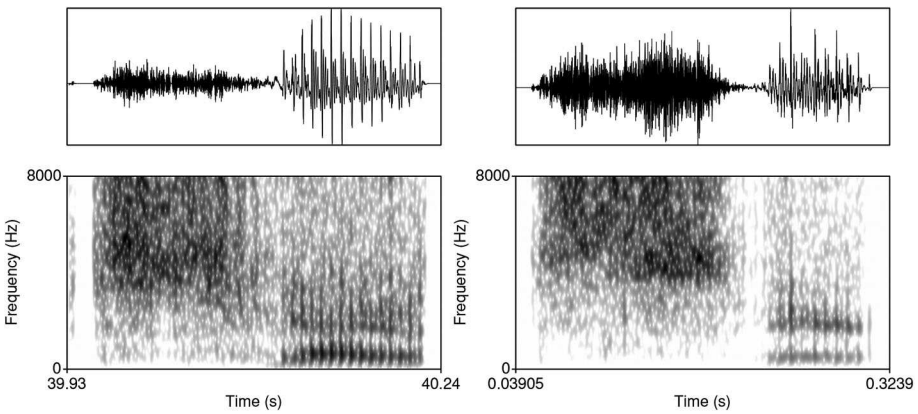

FIGURE 8 | Waveforms and spectrograms (0–8,000 Hz) of two generated outputs of the shape #sV. The sample on the left was generated after 16,715 steps.

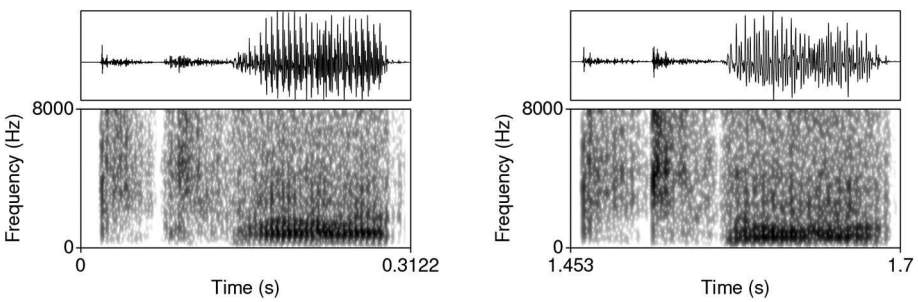

FIGURE 9 | Waveforms and spectrograms (0–8,000 Hz) of two generated outputs of the shape #TTV.

between the network's latent space and generated data based on logistic regression. We first use regression estimates to identify variables with a desired effect on the output by correlating the outputs of the Generator with its corresponding input variables that are uniformly distributed with an interval $(-1, 1)$ during training. We then Generate outputs by setting the identified latent variables to values well beyond the training range (to 4.5, 15, or 25). This method has the potential to reveal the underlying values of latent variables and shed light on the network's internal representations. Using the proposed technique, we can estimate how the network learns to map from latent space to phonetically and phonologically meaningful units in the generated data.

To identify dependencies between the latent space and generated data, we correlate annotations of the output data with

the variables in the latent space (in Section 4.3.1). As a starting point, we choose to identify correlates of the most prominent feature in the training data: the presence or absence of [s]. Any number of other phonetic features can be correlated with this approach (for future directions, see Section 6); applying this technique to other features and other alternations should yield a better understanding of the network's learning mechanisms. Focusing on more than the chosen feature, however, is beyond the scope of this paper.

4.3.1. Regression

First, 3,800 outputs from the Generator network were generated and manually annotated for the presence or absence of [s]. 271 outputs (7.13%) were annotated as involving a segment [s] which is similar to the percentage of data points with [s] in the training data (9.8%). Frication that resembled [s]-like aspiration noise after the alveolar stop and before high vowels was not annotated as including [s][7]. Innovative outputs such as an #[s] without the following vowel or #sV sequences were annotated as including 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 models 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 within the interval $(-1, 1)$ as predictors.

The models explored here do not serve for hypothesis testing, but for exploratory purposes: to identify variables, the effects of which are tested with two independent generative tests (see Sections 4.3.2 and 4.3.3). For this reason, several strategies to reduce the number of variables in the model with different shrinkage techniques 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 (**Figure 11**). 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[8].

TABLE 2 | AIC values of five fitted models with corresponding degrees of freedom (df), fitted with Maximum Likelihood.

	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

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).*

While the number of predictors in the models is high even after shrinkage or exclusion, there is 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 the *vif()* function (in the *car* package; Fox and Weisberg, 2019) is 1.287. All concurvity estimates in the non-linear models 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 in Sections 4.3.2 and 4.3.3, 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 the presence of [s] in generated outputs. **Table 2** lists AIC for each model. The LINEAREXCLUDED model has the lowest AIC score. All six models, however, yield similar results. For further tests based on Lasso regression and Random Forest models that also yield similar results, see Section 3 (**Supplementary Materials**).

To identify the latent variables with the 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 10** plots those values in a descending order. 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 the presence or absence of [s] in output. **Figure 11** plots smooths of the seven predictors (z_5, z_{11}, z_{49}, z_{29}, z_{74}, z_{26}, z_{14}) from a non-linear model SELECT. The smooths show a linear or near-linear relationship between values of the chosen seven variables and the probability of [s] in the output.

Several methods for finding the features that predict the presence or absence of [s] are thus used. Logistic regression is presented here because it is the simplest and easiest to interpret.

[7]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.

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

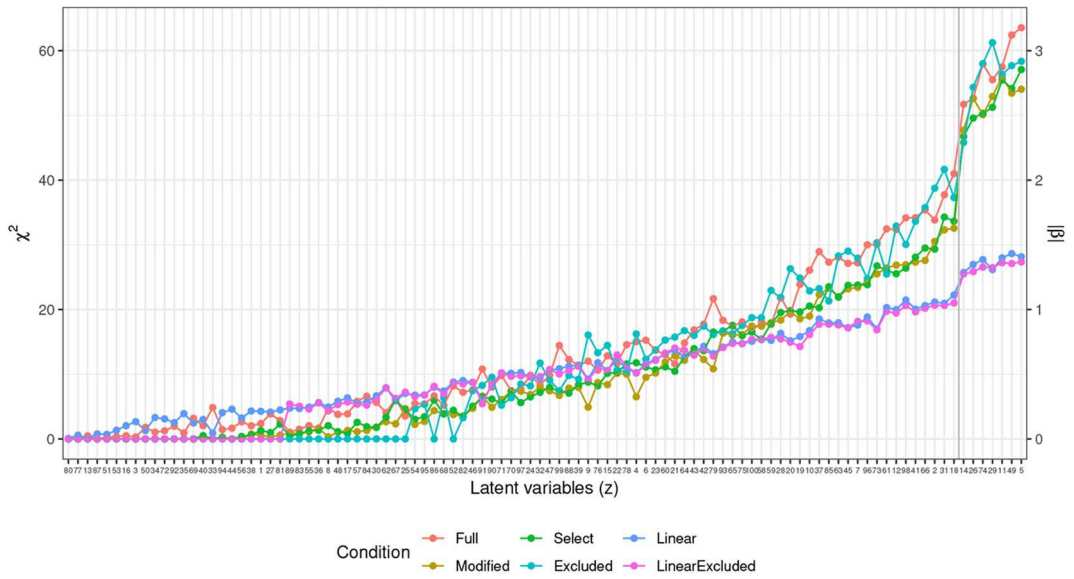

FIGURE 10 | Plot of χ^2 values (left scale) for the 100 predictors across the four generalized additive models. For the two linear models (LINEAR and LINEAR EXCLUDED), estimates of slopes in absolute values ($|\beta|$) are plotted (right scale). The blue vertical line indicates the division between the 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.

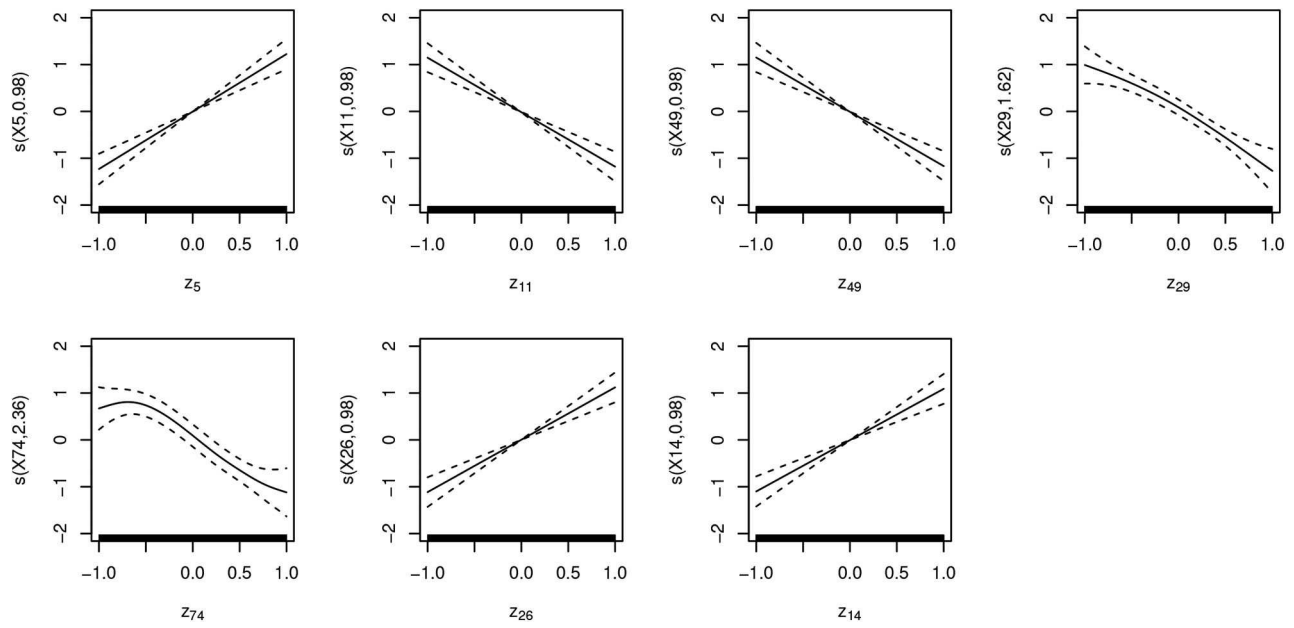

FIGURE 11 | Plots of seven smooth terms with highest χ^2 values in a generalized additive logistic regression model with all 100 latent variables (z) as predictors, estimated with penalty for each term (SELECT). Many of the predictors show linear correlation, which is why a linear logistic regression outputs similar estimates.

In future work, a combination of techniques is recommended to be used for exploratory purposes in a similar way as proposed in this paper. Below, we conduct two independent generative tests to evaluate whether the proposed technique indeed identifies variables that correspond to presence of [s] in the output.

4.3.2. Generative Test 1

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 or negative; see **Figure 11**)

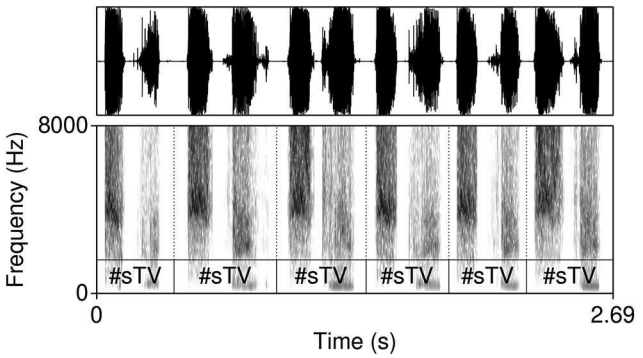

FIGURE 12 | Seven waveforms and spectrograms (0–8,000 Hz) of outputs from the Generator network with the value of z_{11} set at −25. In 96 out of 100 generated samples, the network outputs a sequence containing an [s]. With such a low value of z_{11} (that correlates with amplitude of frication noise), the amplitude of the frication noise reaches the maximum level of 1 in all outputs with [s].

and generated 100 outputs. Altogether seven values in the latent space were thus manipulated, which represents only 7% (7/100) of all latent variables. Of the 100 outputs with manipulated values, 73 outputs included an [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 3,800 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 interval 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 randomly distributed input latent variables [$\chi^2_{(1)} = 792.7, p < 0.0001$]. When z_{11} is −25, the rate goes up to 96 out of 100, also significantly different from random inputs [$\chi^2_{(1)} = 959.8, p < 0.0001$].

4.3.3. Generative Test 2

To further confirm that the regression models identify the variables involved with the presence of [s] in generated outputs, another generative experiment was conducted. In addition to manipulating the seven identified variables, we test the effect of other variables in the latent space on the presence of [s] in the output. If the regression estimates provide reliable information, the variables with higher estimates should have more of an effect on the presence of [s] in the output and vice versa. Testing the entire latent space would be too expensive, which is why we limit our tests to 25 variables with highest estimates from the regression models (which includes the seven chosen variables) and six additional variables with descending regression estimates. Altogether 31/100 variables or 31% of the latent variables are thus analyzed. The variables were chosen in the following way: first, we manipulate values of the first 25 variables with the highest estimates based on regression models in **Figure 10** (7 chosen variables plus additional 18 variables for a total of 25). Because we want to test the effects of the latent variables as evenly as possible and also to test the effects of variables with the lowest regression estimates, we picked 6 additional variables that are distanced from the 25th highest variable in increments of 5 (random choice of variables might miss the variables with lowest estimates).

To perform the generative test of the correlation between the latent space and the proportion of [s] in the output, we set each of the 31 variables at a time to a marginal level well beyond the training interval (to ±4.5), while keeping the rest of the latent space randomly sampled, but constant. In other words, all variables are sampled randomly and held constant across all samples, with the exception of the variable in question at a time that is set to ±4.5. The ±4.5 value was chosen based on manual inspection of generated samples: as is clear from **Figure 15**, changes in amplitude of [s] become increasingly smaller when variables have a value greater than ±3.5. For effects of values beyond 4.5, see **Figure 12**.

One hundred outputs are generated for each of the 31 manipulated latent variables. Altogether 31 × 100 (3,100) outputs were thus analyzed and annotated for the presence or absence of [s] in the output. For example, when the effect of latent variable z_{11} on the proportion of [s] in the output is tested, we set its value to −4.5 while keeping other variables random. One-hundred samples are generated in which the other 99 variables are randomly distributed with the exception of the z_{11} variable (which is set at the marginal level). Samples are annotated for the presence or absence of [s] and the proportion of [s] in the output is calculated from the number of samples with [s] divided by the number of all samples. The same procedure is applied to the other 30 variables examined. To control for the unwanted effects of the latent space on the output, all 99 other variables with the exception of the one manipulated are kept constant across all 31 samples. The 31 data points of this proportion are thus the dependent variable in regression models (**Figure 13**) that test the correlation between the identified variables and [s] in the output.

A beta regression model with the proportion of [s] as the dependent variable and with estimates of the Linear model as the independent variable suggests that there exists a significant linear correlation between the estimates of the regression model and the actual proportion of generated outputs with [s]: $\beta = 1.44, z = 5.07, p < 0.0001$ (for details on model selection, see Section 4, **Supplementary Materials**). In other words, the technique for identifying latent variables that correlate with the presence of [s] in the output based on regression models (in **Figure 10**) successfully identifies such variables. This is confirmed independently: the proportion of generated outputs containing an [s] correlates significantly with its estimates from the regression models.

4.3.4. Interpretation

The regression models in **Figure 10** identify those z-variables in the latent space that have the largest effect on the presence of

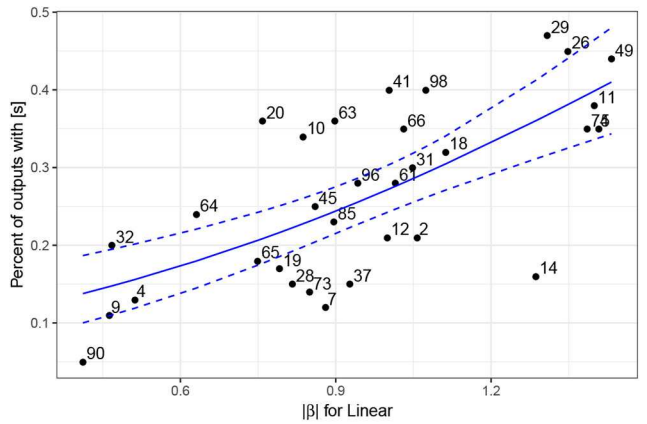

FIGURE 13 | Plot of absolute values of estimates from the Linear model for 31 analyzed latent variables z (numbered on the plot) and the percent of outputs that contain an [s] based on 100 generated samples. The blue solid line represents predicted values based on the beta regression model with estimates of the Linear model as the predictor; the dashed lines represent 95% confidence intervals.

[s] in the output. **Figure 13** confirms that the generator outputs significantly higher proportions of [s] for some variables, while other variables have no effect on the presence of [s]. In other words, variables with lower regression estimates do not affect the proportion of [s] in the output. The proportion of [s] in the output when variables such as z_{90}, z_9, z_4, z_7 are manipulated is very close to the 7.13% of [s] in the output when all z-variables in the latent space are random. It thus appears that the Generator uses portions of the latent space to encode the presence of [s] in the output.

Some latent variables cause a high proportion of [s] in the output despite the regression model estimating their contribution lower than the seven identified latent variables (**Figure 13**) and vice versa. Outputs for variable z_{14} contain frication noise that falls between [s] and [s]-like aspiration, which were difficult to classify (also, the target for [s]-like outputs in this variable is closer to 2.5). The two variables with the highest proportion of [s] in the output that are estimated substantially lower than the seven variables are 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 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 fed to the regression models 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 for these variables. There are only a handful of such variables, and since we are primarily interested in those variables that correspond to [s] both within the training interval and outside of it, we focus our analysis below on the seven variables identified in Section 4.4. The problem with variables in which [s]-outputs are present predominantly outside of the training interval 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 10**) and while several independent generation tests confirm that the seven variables have the strongest effect on the 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 (**Figure 10**) 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 interval (± 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].

4.4. 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 the presence of [s] in the output. As will be shown below, this is not where exploration of the Generator's internal representations should end. We explore whether the mapping between the uniformly distributed input z-variables and the Generator's output signal that resembles speech can be associated with specific phonetic features in that output. The crucial step in this direction is to explore values of the latent space and their phonetic correlates in the output beyond the training interval, i.e., beyond $(-1, 1)$. We observe that the Generator network, while being trained on latent space limited to the interval $(-1, 1)$, learns representations that extend this interval. Even if the input latent variables (z) exceed the training interval, the Generator network outputs samples that closely resemble human speech. Furthermore, the dependencies learned during training extend outside of the $(-1, 1)$ interval. As is argued in Section 4.5, exploring phonetic properties at these marginal values has the potential to 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. Again, the ± 4.5 value was chosen based on manual inspection of generated samples: amplitude of [s] ceases to change substantially past values around ± 3.5 (**Figure 15**). Seven sets of generated samples are thus created, one for each of the seven z values: $z_5, z_{11}, z_{14}, z_{26}, z_{29}, z_{49}$, and z_{74} (with the other 99 z-values randomly sampled, but kept constant for all seven manipulated variables). Each set contains a subset

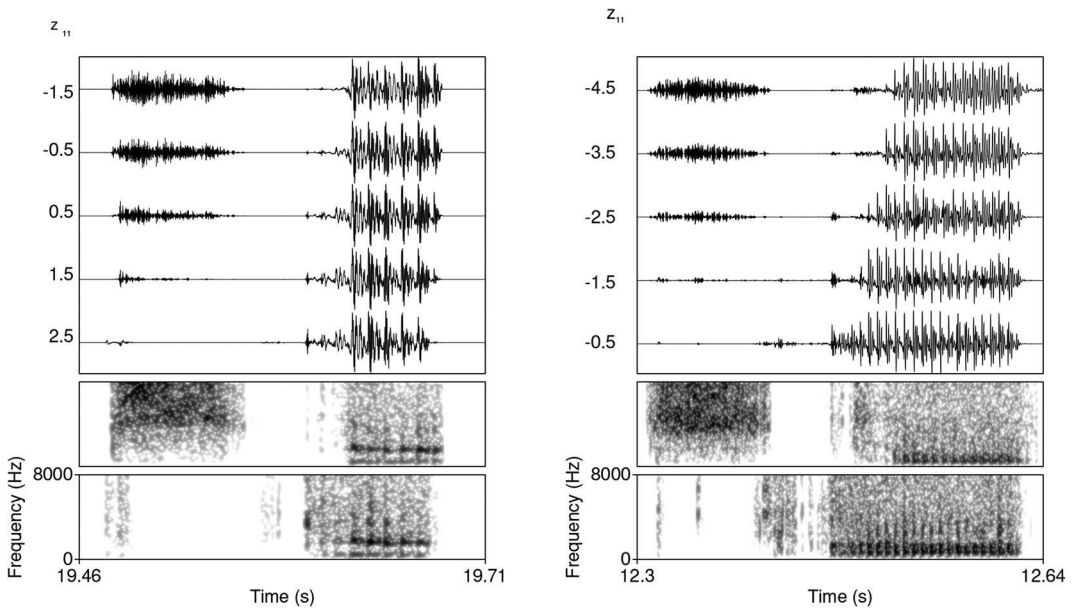

FIGURE 14 | Waveforms and two spectrograms (both $0 - 8,000$ Hz) of generated data with z_{11} variable manipulated and interpolated. The values on the left of waveforms indicate the value of z_{11}. The two spectrograms represent the highest and the lowest value of z_{11}. A clear attenuation of the frication noise is visible until complete disappearance.

of 19 generated outputs that correspond to the interpolated variables from -4.5 to 4.5 in 0.5 increments. Twenty-nine such sets that contained an [s] in at least one set are extracted for analysis (sets that lack an [s] were not analyzed).

A clear pattern emerges in the generated data: the latent variables identified as corresponding to the presence of [s] via regression (**Figure 10**) have direct phonetic correlates and cause changes in amplitude and the 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 interval. 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 14 illustrates this effect. Frication noise of [s] gradually decreases by increasing the value of z_{11} until it completely disappears from the output. 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 appear to be #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 effects. Regardless of these underlying phonetic effects, manipulating the chosen variables has a clear effect of causing [s] to appear in the output and controlling its amplitude.

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 [s] and the following vowel. Outputs gradually change from #sTV to #TV. Only sequences containing an [s] were analyzed; as soon as [s] stops in the output, 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 number of measurements of the interpolated values of z. For each datapoint, maximum intensity of the fricative and maximum intensity of the vowel were extracted in Praat (Boersma and Weenink, 2015) with a 13.3 ms window length (with parabolic interpolation)[9]. **Figure 15** illustrates how manipulating the values of z of the chosen variables from the marginal value ± 4.5 decreases frication noise in the output until [s] is completely absent.

To test whether the decreased frication noise is not part of a general effect of decreased amplitude, we perform significance tests on the ratio of maximum intensity between the frication noise of [s] and the following vowel in the #sTV sequences. **Figure 16** 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 the 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.

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

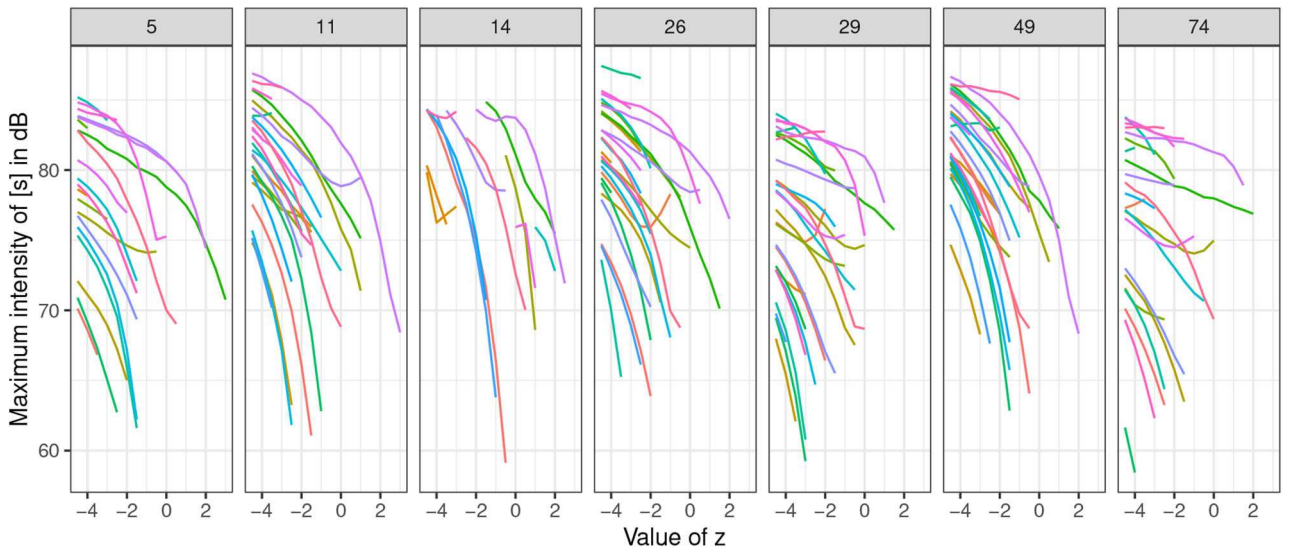

FIGURE 15 | Plots of maximum intensity (in dB) of the fricative part in #sTV sequences when values of the seven *z*-variables are interpolated from the marginal values ±4.5 in 0.5 increments. Each set of generated samples with the randomly sampled latent variables held constant is colored with the same color across the seven *z*-variables. Values of z_5, z_{14}, and z_{26} are inverted for clarity purposes.

The data were fit to a beta regression generalized additive mixed model (in the *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. **Figure 16** 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 S3**, **Supplementary Materials**) and the plots show a clear negative trajectory. In other words, maximum intensity of [s] is increasingly attenuated compared to 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 the presence or absence of [s] in the output; by manipulating the chosen variables to the identified values we can attenuate frication noise of [s] and cause its presence or complete disappearance in the generated data. Again, the discovery of these features is possible because we extend the initial training interval and test predictions on marginal values. In Section 4.5, we analyze further phonetic correlates of each of the seven variables.

4.5. Phonetic Values of Latent Variables

Interpolation of latent variables reveals that the 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 (**Figure 10**) and a decline of outputs with [s] in generated data (**Figure 13**). This suggests that the learning at this stage is gradient and probabilistic rather than fully categorical.

The different latent variables that correspond to the 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) are likely spectral properties of the frication noise. The seven variables are thus similar in the sense that manipulation of their values results in the 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 were analyzed with and extracted with a script from 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 data points. COG, kurtosis, and skew based on power spectra are measured in each of these 1,610 intervals with 750–8,000 Hz Hann 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,

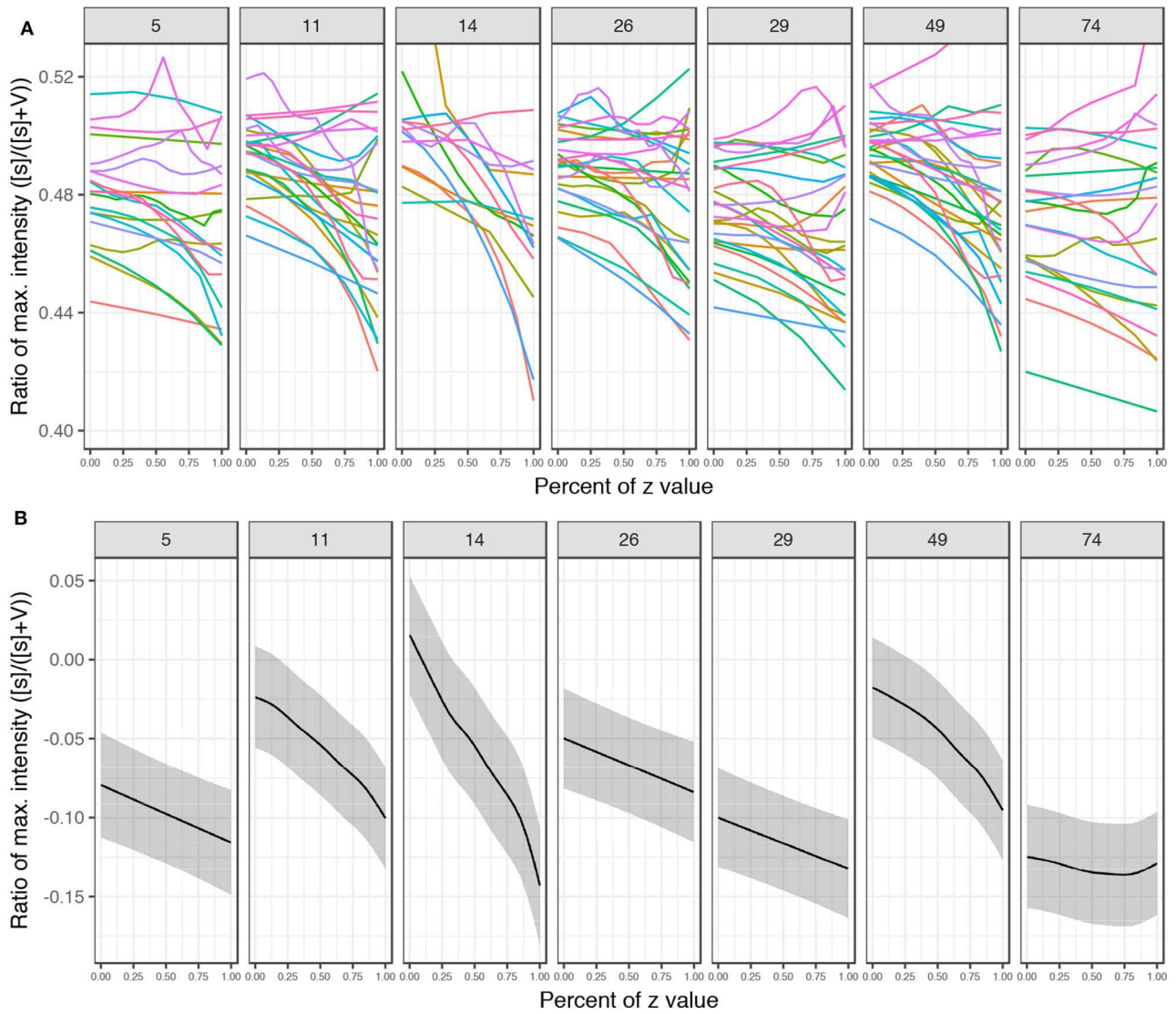

FIGURE 16 | (A) Plots of ratios of maximum intensity between the frication of [s] and phonation of the vowel in #sTV sequences across the seven variables. The interpolated values are normalized where 0 represents the most marginal value of z with [s] in the output and 1 represents the value of z right 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 S3**, **Supplementary Materials**) across the several variables with normalized values.

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 17**). As values of z increase toward the marginal levels (in most cases, ± 4.5), however, clear differentiation in spectral properties emerge between some of the seven z-variables (STRONG; right column in **Figure 17**). 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 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 and non-significant relationship of the six models, see **Tables S4–S9**, **Supplementary Materials**.

In sum, manipulating the latent variables that correspond to [s] in the output not only attenuates frication noise (when vocalic amplitude is controlled for) and causes [s] to surface or disappear from the output, but the different z-variables likely correspond to different phonetic features of the frication noise. At the level before the frication noise ceases from the output, there are no

differences in spectral moments between the latent variables. By setting the values to the marginal levels well beyond the training interval, 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.

5. 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 4.2 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. It is likely, however, that minor local dependencies (such as burst or vowel duration) also contribute to this distribution. While it is currently not possible to disambiguate between the local and non-local effects, it is desirable for a model to capture all possible dependencies, as speech production and perception often employ several cues as well.

5.1. Parallels in Human Behavior

Several similarities emerge between the training of the Generative Adversarial networks and L1 acquisition. The training data in the GAN model is of course not completely naturalistic (even though the inputs are raw audio recordings of speech data): the network is trained on only a subset of sound sequences that a language learning infant is exposed to. 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. Such comparisons have both the potential to inform computational models of human cognition and conversely, shed light on the question of how neural networks learn the data.

While the generated outputs contain evidence that the network learns the conditional distribution of VOT duration, a proportion of outputs violates this distribution. In fact, in approximately 12.2% of the #sTV sequences, the Generator outputs VOT durations that are longer than any VOT duration in the #sTV condition in the training data. This suggests that the model learns the conditional distribution, but that the learning is imperfect and the Generator occasionally violates the distribution. Crucially, these outputs that violate the training data closely 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 exists 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 violates the allophonic distribution and is longer than any VOT in the #TV condition.

Further parallels exist between the Generator's behavior and L2 acquisition and speech errors. Studies on L2 acquisition of VOT durations in #sTV and #TV sequences suggest that learners start with a smaller distinction between the two groups and acquire the non-aspiration rule after [s] only with more exposure (Haraguchi, 2003). A smaller initial difference between the two conditions in L2 acquisition, for example, increases from Japanese learners of English with little exposure when compared with learners with more exposure (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, which is parallel to the Generator network outputting a long VOT in the #sTV condition that violate the training data distributions.

Finally, the Generator network violating the VOT distribution resembles the behavior of patients with speech impairments. Buchwald and Miozzo (2012) analyzed VOT durations of two patients with apraxia of speech that present cluster production errors, i.e., clusters of the structure #sTV are simplified to #TV. One patient outputs long VOT durations in the #sTV condition (after the cluster is simplified). VOT durations in the #sTV clusters in this patient correspond to VOT durations of singleton stops (#TV). The other patient also simplifies the cluster, but outputs shorter VOT durations in the #sTV condition, maintaining the underlying distribution. It is hypothesized that the first patient (with long VOT durations in the #sTV condition) shows signs of impairment that operates on the phonological level: because phonological

computation is impaired, the patient fails to output shorter VOT durations in the #sTV condition. In other words, there are no motor planning mechanisms that would prevent the patient from producing shorter VOT durations in the #sTV condition, which is why the error is assumed to operate on the phonological level — a phonological rule fails to apply, which results in long VOT in the #sTV condition. The second patient, on the other hand, is hypothesized to shows traces of phonetic execution impairment, while the phonological computation (short VOT in the #sTV condition) is intact. The outputs of the Generator network that violate the training data are parallel to the behavior of the patient with assumed phonological impairment: in 12.2% of cases, the network outputs long VOT duration in the #sTV condition that is longer than any VOT duration in the same condition in the training data. Since the network lacks any articulatory component (see also discussion below), motor planning factors cannot explain the Generator's violations of the distributions in the training data.

As indicated by examples in **Figures 8**, **9**, the network also generates segmentally innovative outputs for which no evidence was available in the training data. A subset of the innovative outputs, such as #sV and #TTV sequences, are consistent with linguistic behavior in humans. The Generator's innovative outputs thus closely resemble one of the main properties 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 treat 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] (which results in the #TV sequence), deletion of the stop is robustly attested as well in L1 acquisition (resulting in #sV), both in the general population and in children with speech disorders (Catts and Kamhi, 1984; Ohala, 1999; Gerlach, 2010; Syrika et al., 2011). While this deletion likely involves articulatory factors that are lacking in our model, the fact that segmental units can be deleted from the output and recombined in L1 acquisition resembles the deletion in the Generator's innovative outputs, such as the #sV sequence.

These innovative outputs of the Generator's network have potential for contributing to our understanding of the evolution of phonology in language evolution in general (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 traces of this behavior: in addition to learning to reproduce the input, it learns to recombine segments into novel and unobserved sequences. The exact details of modeling phonological evolution with Generative Adversarial architecture is, however, beyond the scope of the present paper.

5.2. Latent Variables as Correlates of Features

In Section 4.3, we propose 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. This allows for an in-depth analysis of phonetic effects of each latent variable in the generated data. Regression models identify those variables in the latent space 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 also uses latent space to encode what would be an approximate equivalent of phonological representations in the broadest sense—absence or presence of a segment.

While the presence of [s] in the output is controlled by multiple latent variables, each of the variables likely 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 [s] ceases from the output, a clear differentiation emerges when the values are set to the marginal level (**Figure 17**). The seven variables thus likely have a phonetic function: controlling various spectral properties of the frication noise.

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 with neural networks has recently seen success in the autoencoder architecture (Räsänen et al., 2016; Eloff et al., 2019; Shain and Elsner, 2019). 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 imperfectly correspond to phonological features as posited by phonological theory. As was argued in Section 4.3, our model shows traces of imperfect self-organizing of phonetic features (e.g., spectral moments) and phonological representations (e.g., the presence of [s]) 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 entire phonological feature matrices (such as the full presence of [s] in the output) and phonetic

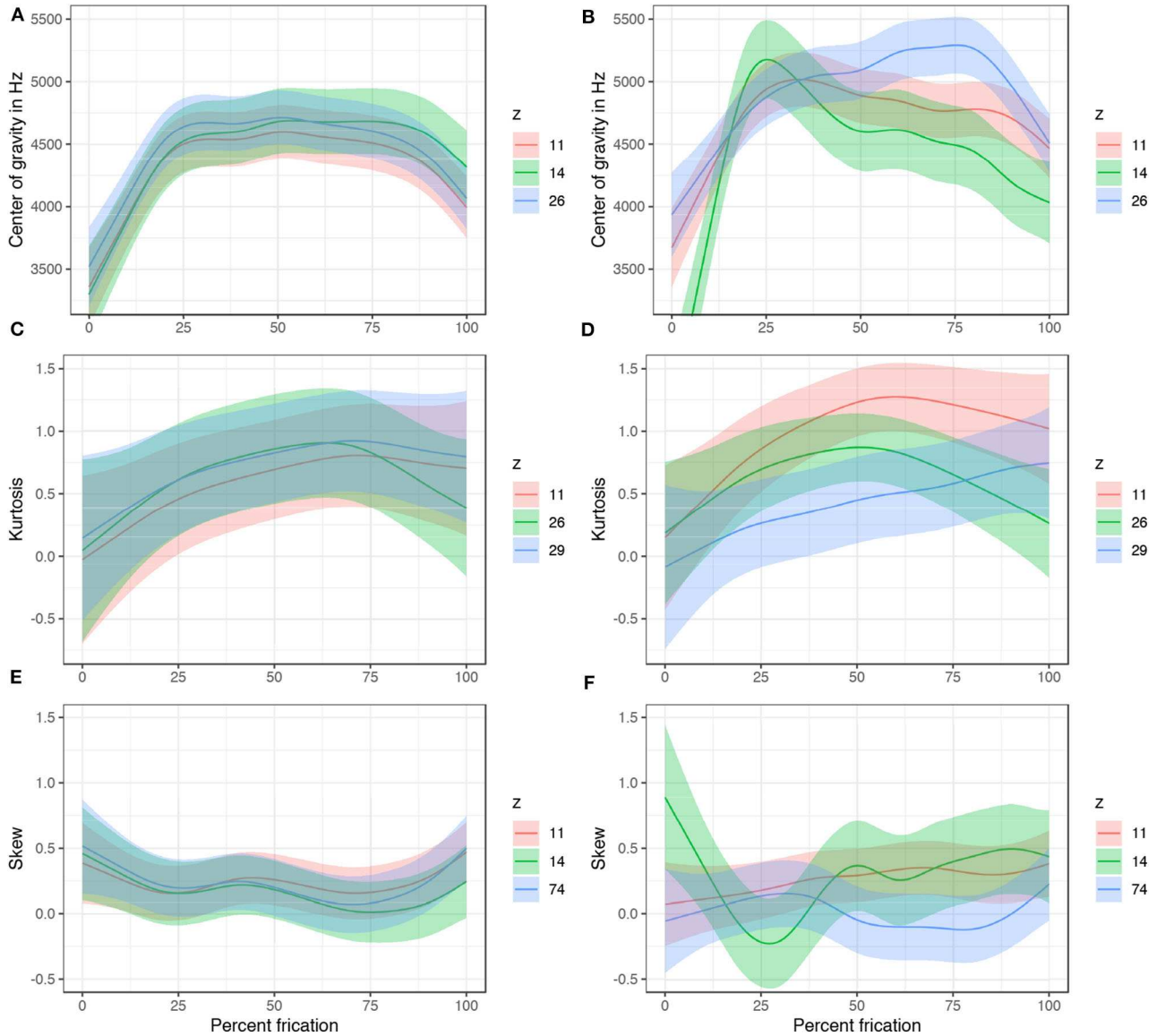

FIGURE 17 | A subset of predicted values of COG, kurtosis, and skew with 95% CIs in two conditions: WEAK with *z*-variables at the value before [s] ceases from the output (left column) and STRONG (right column) with the most marginal value with [s]-output (±4.5 in most cases). Predicted values are based on generalized additive models in **Tables S4–S9** (**Supplementary Materials**). 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 S1**, **Supplementary Materials**.

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 actual phonological features such as [±sonorant] emerge in the training. This task is left for future work. Despite these differences, the fact that we can actively control the 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 representations.

On a very speculative level, the latent space of the Generator's network might have a conceptual 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 planning with some actual clinical

applications that result from such models[10]. Recently, high-density direct cortical surface (electrocorticographic) recordings of the superior temporal gyrus during open brain surgery in Mesgarani et al. (2014) suggest 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.

5.3. Future Directions

Among the objections against modeling phonological learning with Generative Adversarial Networks might be that the model is too powerful and that it overgenerates. First, it has been shown in numerous examples that phonology, while being computationally limited (Heinz, 2011; Avcu et al., 2017), is more powerful than the attested phonological typology. Subjects in the artificial grammar learning paradigm are, for example, able to successfully learn alternations that never surface in natural languages (Glewwe, 2017; Glewwe et al., 2017; Avcu, 2018; Beguš, 2018a,b). Second, 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 a more serious shortcoming: a model of phonology should at minimum derive the observed phonological processes. Finally, the main reason the proposed 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 would be problematic if the goal of the current model were a network that models phonetic and phonological learning both on the articulatory and the cognitive levels. 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 for computational models of the cognitive basis of speech production and perception. It is likely that speech acquisition involves various different types of learning. Learning of motor-planning on the articulatory level is likely different from learning of articulatory targets based on perception, which is in turn likely controlled by other systems than learning of abstract symbol manipulation on the phonological level, even though these levels are interconnected in acquisition. Among the evidence that exemplifies the different levels of representation 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 and therefore only influenced by the neural network architecture. 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).

While the proposal in this paper does not directly address the discussion between generative and exemplar-based approaches to phonology, the GAN models have the potential to offer some insights into this discussion as well. The results of the computational experiments suggest that the network learns to output data consistent with the training data without grammar-specific assumptions, which would support the exemplar-based approaches to phonology. On the other hand, the Generator network does seem to compress phonological information in its latent space in a way that does not directly correspond to stored exemplars. Further explorations of the latent space should shed light on this long-standing discussion.

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 significant losses in the signal. A GAN trained on spectral transformations would likely be closer to reality, as human auditory mechanisms resemble 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.

6. CONCLUSION

The results of this paper suggest that we can model phonology not only with rules (as in rule-based approaches; Chomsky and Halle, 1968), exemplars (Pierrehumbert, 2001), finite-state automata (Heinz, 2010; Chandlee, 2014), input-output optimization (as in Optimality Theory; Prince and Smolensky, 2004), or with neural network architecture that already assumes some level of abstraction (see Section 1), but as a mapping between random latent variables and output data in deep neural networks that are trained in an unsupervised manner from raw acoustic data. To the author's knowledge, this is the first paper testing learning of allophonic distributions in an unsupervised manner from raw acoustic data using neural networks and the first proposal to use GANs for modeling language acquisition. The Generative Adversarial model of phonology (trained on an implementation

[10]Warlaumont and Finnegan (2016) propose a model of infant babbling that involves spiking neural networks and speech synthesis. While the model does not take any speech as an input, babbling emerges even if the objective for the simulation is maximization of perceptual salience.

of DCGAN architecture for audio data in Donahue et al., 2019) derives outputs that resemble speech from latent variables. The results of the computational experiment suggest that the network learns the conditional allophonic distribution of VOT duration. We propose a technique that identifies variables in the latent space that correspond to phonetic and phonological properties in the output, such as the presence of [s], and show that by manipulating these values, we can generate data with or without [s] in the output as well as control its intensity and spectral properties of its frication noise. While at least seven latent variables control the presence of [s], each of them likely has a phonetic function that controls spectral properties of the frication noise. The proposed technique thus suggests that the Generator network learns to encode phonetic and phonological information in its latent space. Finally, the model generates innovative outputs, suggesting its productive nature. The behavior of the model is compared against speech acquisition, speech errors, and speech impairment; several parallels are identified.

The current proposal models one allophonic distribution in English. Training GAN networks on further processes and on languages other than English as well as probing the networks at different training steps 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. This paper outlines a methodology for establishing internal representations and testing predictions against generated data, but represents just a first step in a broader task of modeling phonetic and phonological learning in a Generative Adversarial framework.

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

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This research was funded by a grant to new faculty at the University of Washington. Publication made possible in part by support from the Berkeley Research Impact Initative (BRII) sponsored by the UC Berkeley Library.

ACKNOWLEDGMENTS

I would like to thank Sameer Arshad for slicing data from the TIMIT database and Heather Morrison for annotating data. Parts of this research were published in Beguš (2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00044/full#supplementary-material

REFERENCES

Abramson, A. S., and Whalen, D. (2017). Voice onset time (VOT) at 50: theoretical and practical issues in measuring voicing distinctions. *J. Phonet.* 63, 75–86. doi: 10.1016/j.wocn.2017.05.002

Adlam, B., Weill, C., and Kapoor, A. (2019). Investigating under and overfitting in wasserstein generative adversarial networks. *arXiv [Preprint]*. arXiv:1910.14137.

Alanazi, S. (2018). *The acquisition of English stops by Saudi L2 learners* (Ph.D. thesis). University of Essex, Essex, United Kingdom.

Alderete, J., and Tupper, P. (2018a). "Connectionist approaches to generative phonology," in *The Routledge Handbook of Phonological Theory*, eds A. Bosch and S. J. Hannahs (New York, NY: Routledge), 360–390. doi: 10.4324/9781315675428-13

Alderete, J., and Tupper, P. (2018b). Phonological regularity, perceptual biases, and the role of phonotactics in speech error analysis. *Wiley Interdiscipl. Rev. Cogn. Sci.* 9:e1466. doi: 10.1002/wcs.1466

Alderete, J., Tupper, P., and Frisch, S. A. (2013). Phonological constraint induction in a connectionist network: learning ocp-place constraints from data. *Lang. Sci.* 37, 52–69. doi: 10.1016/j.langsci.2012.10.002

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, eds. D. Precup and Y. W. Teh (Sydney, NSW: PMLR; International Convention Centre), vol. 70 of Proceedings of Machine Learning Research, 214-223

Avcu, E. (2018). "Experimental investigation of the subregular hypothesis," in *Proceedings of the 35th West Coast Conference on Formal Linguistics*, eds W. G. Bennett, L. Hracs, and D. R. Storoshenko (Somerville, MA: Cascadilla), 77–86.

Avcu, E., Shibata, C., and Heinz, J. (2017). "Subregular complexity and deep learning," in *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017)*, eds S. Dobnik and S. Lappin. (Gothenburg), 20–33.

Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: some solutions offered by Generalized Additive Mixed Models. *arXiv [Preprint]*. arXiv:1601.02043.

Barlow, J. A. (2001). Case study. *Lang. Speech Hear. Serv. Sch.* 32, 242–256. doi: 10.1044/0161-1461(2001/022)

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science* 364:6439. doi: 10.1126/science.aav9436

Beguš, G. (2018a). Post-nasal devoicing and the blurring process. *J. Linguist.* 55, 689–753. doi: 10.1017/S00222267180 0049X

Beguš, G. (2018b). *Unnatural phonology: a synchrony-diachrony interface approach* (Ph.D. thesis). Harvard University, Cambridge, MA, United States.

Beguš, G. (2020). "Modeling unsupervised phonetic and phonological learning in Generative Adversarial Phonology," in *Proceedings of the*

Society for Computation in Linguistics: Vol. 3 (New Orleans, LA), 15. doi: 10.7275/nbrf-1a27

Boersma, P., and Weenink, D. (2015). *PRAAT: Doing Phonetics by Computer [Computer Program]*. version 5.4.06. Available online at: http://www.praat.org/ (Retrieved February 21, 2015).

Bond, Z. S. (1981). A note concerning /s/ plus stop clusters in the speech of language-delayed children. *Appl. Psycholinguist.* 2, 55–63. doi: 10.1017/S0142716400000655

Bond, Z. S., and Wilson, H. F. (1980). /s/ plus stop clusters in children's speech. *Phonetica* 37, 149–158. doi: 10.1159/000259988

Buchwald, A., and Miozzo, M. (2012). Phonological and motor errors in individuals with acquired sound production impairment. *J. Speech Lang. Hear. Res.* 55, S1573–S1586. doi: 10.1044/1092-4388(2012/11-0200)

Bybee, J. (1999). "Usage-based phonology," in *Functionalism and Formalism in Linguistics*, Vol. 1, eds M. Darnell, E. Moravcsik, F. Newmeyer, M. Noonan, and K. Wheatley (Amsterdam: John Benjamins), 211–242.

Catts, H. W., and Jensen, P. J. (1983). Speech timing of phonologically disordered children. *J. Speech Lang. Hear. Res.* 26, 501–510. doi: 10.1044/jshr.2604.501

Catts, H. W., and Kamhi, A. G. (1984). Simplification of /s/ + stop consonant clusters. *J. Speech Lang. Hear. Res.* 27, 556–561. doi: 10.1044/jshr.2704.556

Chandlee, J. (2014). *Strictly local phonological processes* (Ph.D. thesis). University of Delaware, Newark, DE, United States.

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English*. New York, NY: Harper & Row.

Clements, G. N. (1985). The geometry of phonological features. *Phonol. Yearbook* 2, 225–252. doi: 10.1017/S0952675700000440

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

Davis, S., and Cho, M.-H. (2006). The distribution of aspirated stops and /h/ in American English and Korean: an alignment approach with typological implications. *Linguistics.* 41, 607–652. doi: 10.1515/ling.2003.020

de Boer, B. (2000). Self-organization in vowel systems. *J. Phonet.* 28, 441–465. doi: 10.1006/jpho.2000.0125

de Lacy, P. (2006). Transmissibility and the role of the phonological component: a theoretical synopsis of evolutionary phonology. *Theor. Linguist.* 32, 185–196. doi: 10.1515/TL.2006.012

de Lacy, P., and Kingston, J. (2013). Synchronic explanation. *Nat. Lang. Linguist. Theory* 31, 287–355. doi: 10.1007/s11049-013-9191-y

Donahue, C., Balsubramani, A., McAuley, J. J., and Lipton, Z. C. (2017). Semantically decomposing the latent spaces of generative adversarial networks. *CoRR arXiv [preprint]. arXiv*:1705.07904.

Donahue, C., McAuley, J. J., and Puckette, M. S. (2019). "Adversarial audio synthesis," in *7th International Conference on Learning Representations, ICLR 2019* (New Orleans, LA: OpenReview.net) Available online at: https://arxiv.org/abs/1802.04208

Dresher, B. E. (2015). The motivation for contrastive feature hierarchies in phonology. *Linguist. Variat.* 15, 1–40. doi: 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. doi: 10.1016/j.cognition.2017.11.008

Eloff, R., Nortje, A., van Niekerk, B., Govender, A., Nortje, L., Pretorius, A., et al. (2019). "Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks," in *Proc. Interspeech 2019* (Graz), 1103–1107. doi: 10.21437/Interspeech.2019-1518

Ernestus, M. (2011). "Gradience and categoricality in phonological theory," in *The Blackwell Companion to Phonology*, eds. M. van Oostendorp, C. j. Ewen, E. Hume, and K. Rice (Malden, MA: Wiley Blackwell), 1–22. doi: 10.1002/9781444335262.wbctp0089

Faruqui, M., Tsvetkov, Y., Neubig, G., and Dyer, C. (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* (San Diego, CA: Association for Computational Linguistics), 634–643. doi: 10.18653/v1/N16-1077

Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression, 3rd Edn*. Thousand Oaks CA: Sage.

Fruehwald, J. (2016). The early influence of phonology on a phonetic change. *Language* 92, 376–410. doi: 10.1353/lan.2016.0041

Fruehwald, J. (2017). The role of phonology in phonetic change. *Annu. Rev. Linguist.* 3, 25–42. doi: 10.1146/annurev-linguistics-011516-034101

Futrell, R., Albright, A., Graff, P., and O'Donnell, T. J. (2017). A generative model of phonotactics. *Trans. Assoc. Comput. Linguist.* 5, 73–86. doi: 10.1162/tacl_a_00047

Gahl, S., and Yu, A. C. L. (2006). Introduction to the special issue on exemplar-based models in linguistics. *Linguist. Rev.* 23, 213–216. doi: 10.1515/TLR.2006.007

Garofolo, J. S., Lamel, L., Fisher, M. W., Fiscus, J., Pallett, S. D., Dahlgren, L. N., et al. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download*. Philadelphia: Linguistic Data Consortium.

Gaskell, M., Hare, M., and Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cogn. Sci.* 19, 407–439. doi: 10.1207/s15516709cog1904_1

Gerlach, S. R. (2010). *The acquisition of consonant feature sequences: harmony, metathesis and deletion patterns in phonological development* (Ph.D. thesis). University of Minnesota, Minneapolis, MN, United States.

Gibson, K. R., Tallerman, M., and MacNeilage, P. F. (2012). "The evolution of phonology," in *The Oxford Handbook of Language Evolution*, eds K. R. Gibson and M. Tallerman (Oxford: Oxford University Press). doi: 10.1093/oxfordhb/9780199541119.001.0001

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).

Glewwe, E., Zymet, J., Adams, J., Jacobson, R., Yates, A., Zeng, A., et al. (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).

Goldwater, S., and Johnson, M. (2003). "Learning OT constraint rankings using a maximum entropy model," in *Proceedings of the Workshop on Variation within Optimality Theory*, eds J. Spenader, A. Eriksson, and O. Dahl (Stockholm: Stockholm University), 111–120.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Red Hook, NY: Curran Associates, Inc.), 2672–2680.

Guenther, F. H. (2016). *Neural Control of Speech*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/10471.001.0001

Guenther, F. H., and Vladusich, T. (2012). A neural theory of speech acquisition and production. *J. Neurolinguist.* 25, 408–422. doi: 10.1016/j.jneuroling.2009.08.006

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Red Hook, NY: Curran Associates, Inc.), 5767–5777.

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* (Okayama), 83-91.

Hayes, B. (1999). "Phonetically-driven phonology: the role of optimality theory and inductive grounding," in *Functionalism and Formalism in Linguistics, Volume I: General Papers*, eds M. Darnell and E. Moravscik (Amsterdam: John Benjamins), 243–285. doi: 10.1075/slcs.41.13hay

Hayes, B., and White, J. (2013). Phonological naturalness and phonotactic learning. *Linguist. Inq.* 44, 45–75. doi: 10.1162/LING_a_00119

Hayes, B., and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguist. Inq.* 39, 379–440. doi: 10.1162/ling.2008.39.3.379

Heinz, J. (2010). Learning long-distance phonotactics. *Linguist. Inq.* 41, 623–661. doi: 10.1162/LING_a_00015

Heinz, J. (2011). Computational phonology–part II: grammars, learning, and the future. *Lang. Linguist. Compass* 5, 153–168. doi: 10.1111/j.1749-818X.2011.00268.x

Inkelas, S., and Shih, S. S. (2017). "Looking into segments," in *Proceedings of the Forty-Fifth Annual Meeting of the North East Linguistic Society*, eds K. Jesney, C. O'Hara, C. Smith, and R. Walker (Washington, DC: Linguistic Society of America), 1–18. doi: 10.3765/amp.v4i0.3996

Iverson, G. K., and Salmons, J. C. (1995). Aspiration and laryngeal representation in Germanic. *Phonology* 12, 369–396. doi: 10.1017/S0952675700002566

Jarosz, G. (2019). Computational modeling of phonological learning. *Annu. Rev. Linguist.* 5, 67–90. doi: 10.1146/annurev-linguistics-011718-011832

Johnson, K. (1997). "Speech perception without speaker normalization: an exemplar model," in *Talker Variability in Speech Processing* (San Diego, CA: Academic Press), 145–165.

Johnson, K. (2007). "Decisions and mechanisms in exemplar-based phonology," in *Experimental Approaches to Phonology*, eds M. J. Solé, P. S. Beddor, and M. Ohala (Oxford: Oxford University Press), 25–40.

Kamper, H., Elsner, M., Jansen, A., and Goldwater, S. (2015). "Unsupervised neural network based feature extraction using weak top-down constraints," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brisbane, QLD), 5818–5822. doi: 10.1109/ICASSP.2015.7179087

Kaplan, A. (2017). "Exemplar-based models in linguistics," in *Oxford Bibliographies in Linguistics*, ed M. Aronoff (Oxford: Oxford University Press). doi: 10.1093/obo/9780199772810-0201

Kello, C., and Plaut, D. (2003). "The interplay of perception and production in phonological development: beginnings of a connectionist model trained on real speech," in *5th International Congress of Phonetic Sciences*, eds M. J. Solé, D. Recasens, and J. Romero (Barcelona), pages 297–300.

Keyser, S. J., and Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language* 82, 33–63. doi: 10.1353/lan.2006.0051

Kingston, J., and Diehl, R. L. (1994). Phonetic knowledge. *Language* 70, 419–454. doi: 10.1353/lan.1994.0023

Kirby, J., and Sonderegger, M. (2015). Bias and population structure in the actuation of sound change. *arXiv [preprint]*. arXiv:1507.04420.

Kirov, C. (2017). "Recurrent neural networks as a strong baseline for morphophonological learning," *Poster Presented at 2017 Meeting of the Linguistic Society of America* (Austin, TX). Available online at: https://ckirov.github.io/papers/lsa2017.pdf (accessed October 7, 2019).

Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron* 67, 713–727. doi: 10.1016/j.neuron.2010.08.038

Lee, C.-y., and Glass, J. (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)* (Jeju Island: Association for Computational Linguistics), 40–49.

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

Legendre, G., Sorace, A., and Smolensky, P. (2006). "The optimality theory–harmonic grammar connection," in *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, eds P. Smolensky and G. Legendre (Cambridge, MA: MIT Press), 339–402.

Lennes, M. (2003). *f0-f1-f2-Intensity_PRAAT_Script*. PRAAT script. Modified by Dan McCloy, Esther Le Grésauze, and Gašper Beguš.

Lillicrap, T. P., and Kording, K. P. (2019). What does it mean to understand a neural network? *arXiv [preprint]*. arXiv:1907.06374.

Lipton, Z. C., and Tripathi, S. (2017). Precise recovery of latent vectors from generative adversarial networks. *CoRR arXiv [preprint]*. arXiv:1702.04782.

Lisker, L. (1984). How is the aspiration of English /p, t, k/ "predictable"? *Lang. Speech* 27, 391–394. doi: 10.1177/002383098402700409

Lisker, L., and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 384–422. doi: 10.1080/00437956.1964.11659830

Lowenstein, J. H., and Nittrouer, S. (2008). Patterns of acquisition of native voice onset time in English-learning children. *J. Acous. Soc. Am.* 124, 1180–1191. doi: 10.1121/1.2945118

Macken, M. A., and Barton, D. (1980). The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *J. Child Lang.* 7, 41–74. doi: 10.1017/S0305000900007029

Macken, M. A., and Ferguson, C. A. (1981). Phonological universals in language acquisition*. *Ann. N. Y. Acad. Sci.* 379, 110–129. doi: 10.1111/j.1749-6632.1981.tb42002.x

Mahalunkar, A., and Kelleher, J. D. (2018). "Using regular languages to explore the representational capacity of recurrent neural architectures," in *Artificial Neural Networks and Machine Learning-ICANN 2018*, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis (Cham: Springer International Publishing), 189–198. doi: 10.1007/978-3-030-01424-7_19

Martin, A., Peperkamp, S., and Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cogn. Sci.* 37, 103–124. doi: 10.1111/j.1551-6709.2012.01267.x

McClelland, J. L., and Elman, J. L. (1986). The trace model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McLeod, S., van Doorn, J., and 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* (Adelaide), 331–336.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology* 25, 83–127. doi: 10.1017/S0952675708001413

Moreton, E., and Pater, J. (2012a). Structure and substance in artificial-phonology learning. Part I, Structure. *Lang. Linguist. Compass* 6, 686–701. doi: 10.1002/lnc3.363

Moreton, E., and Pater, J. (2012b). Structure and substance in artificial-phonology learning. Part II, Substance. *Lang. Linguist. Compass* 6, 702–718. doi: 10.1002/lnc3.366

Nguyen, N., and Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. *J. Phonet.* 53, 46–54. doi: 10.1016/j.wocn.2015.08.004

Ohala, D. K. (1999). The influence of sonority on children's cluster reductions. *J. Commun. Disord.* 32, 397–422. doi: 10.1016/S0021-9924(99)00018-0

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 (Berlin: Springer), 1171–1176. doi: 10.1007/3-540-44668-0_163

Oudeyer, P.-Y. (2002). "Phonemic coding might result from sensory-motor coupling dynamics," in *From animals to animats 7: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, eds B. Hallam, D. Floreano, J. Hallam, G. Hayes, and J.-A. M. Hallam (Cambridge, MA: MIT Press), 406–416.

Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *J. Theor. Biol.* 233, 435–449. doi: 10.1016/j.jtbi.2004.10.025

Oudeyer, P.-Y. (2006). *Self-Organization in the Evolution of Speech*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199289158.001.0001

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Pater, J. (2009). Weighted constraints in generative linguistics. *Cogn. Sci.* 33, 999–1035. doi: 10.1111/j.1551-6709.2009.01047.x

Pater, J. (2019). Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Language*. 95. doi: 10.1353/lan.2019.0005

Pierrehumbert, J. (2001). "Exemplar dynamics: word frequency, lenition, and contrast," in *Frequency Effects and the Emergence of Lexical Structure*, J. L. Bybee and P. J. Hopper (Amsterdam: John Benjamins), 137–157. doi: 10.1075/tsl.45.08pie

Plaut, D. C. and Kello, C. T. (1999). "The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach," in *The Emergence of Language*, ed. B. MacWhinney (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 381–415.

Pouplier, M., Marin, S., and Waltl, S. (2014). Voice onset time in consonant cluster errors: can phonetic accommodation differentiate cognitive from motor errors? *J. Speech Lang. Hear. Res.* 57, 1577–1588. doi: 10.1044/2014_JSLHR-S-12-0412

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

Prince, A., and Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Malden, MA. First published as Tech. Rep. 2, Rutgers University Center for Cognitive Science.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv [preprint]*. arXiv:1511.06434.

Räsänen, O., Nagamine, T., and Mesgarani, N. (2016). "Analyzing distributional learning of phonemic categories in unsupervised deep neural networks," in *Procedings of the 38th Annual Conference of the Cognitive Science Society*, eds A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell (Austin, TX: Cognitive Science Society). Available online at: https://cogsci.mindmodeling.org/2016/papers/0308/paper0308.pdf

Rawski, J., and Heinz, J. (2019). No free lunch in linguistics or machine learning: response to pater. *Language* 95, e125–e135. doi: 10.1353/lan.2019.0004

Rentz, B. (2017). spectral_moments.praat. praat script. Available online at: https://github.com/rentzb/praat-scripts/blob/master/spectral_moments.praat

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Saffran, J. R., Werker, J. F., and Werner, L. A. (2007). "The infant's auditory world: hearing, speech, and the beginnings of language," in *Handbook of Child Psychology*, eds. W. Damon and R. M. Lerner (Hoboken, NJ: Wiley). doi: 10.1002/9780470147658.chpsy0202

Schatz, T., Feldman, N., Goldwater, S., Cao, X. N., and Dupoux, E. (2019). Early phonetic learning without phonetic categories - insights from machine learning. *PsyArXiv*. doi: 10.31234/osf.io/fc4wh.

Shain, C., and Elsner, M. (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)*, 69–85 (Minneapolis, MN: Association for Computational Linguistics). doi: 10.18653/v1/N19-1007

Silfverberg, M. P., Mao, L., and Hulden, M. (2018). "Sound analogies with phoneme embeddings," in *Proceedings of the Society for Computation in Linguistics (SCiL) 2018* (Salt Lake City, UT), 136–144. doi: 10.7275/R5NZ85VD

Silverman, D. (2017). *A Critical Introduction to Phonology: Functional and Usage-Based Perspectives*. London: Bloomsbury Publishing.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv [preprint]*. arXiv:1703.05339.

Syrika, A., Nicolaidis, K., Edwards, J., and Beckman, M. E. (2011). Acquisition of initial /s/-stop and stop-/s/ sequences in Greek. *Lang. Speech* 54, 361–386. PMID: 22070044. doi: 10.1177/0023830911402597

Thiolliére, R., Dunbar, E., Synnaeve, G., Versteegh, M., and Dupoux, E. (2015). "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proceedings of Interspeech* (Dresden).

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Travaux de Cercle linguistique de Prague, Prague.

Vaux, B. (2002). *Aspiration in English* (Ms.). Harvard University, Cambridge, MA, United Kingdom.

Vaux, B., and Samuels, B. (2005). Laryngeal markedness and aspiration. *Phonology* 22, 395–436. doi: 10.1017/S0952675705000667

Warlaumont, A. S., and Finnegan, M. K. (2016). Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PLoS ONE* 11:e0145096. doi: 10.1371/journal.pone.0145096

Weber, N., Shekhar, L., and 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* (New Orleans, LA: Association for Computational Linguistics), 24–27. doi: 10.18653/v1/W18-1004

Wedel, A. (2006). Exemplar models, evolution and language change. *Linguist. Rev*. 23, 247–274. doi: 10.1515/TLR.2006.010

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition* 130, 96–115. doi: 10.1016/j.cognition.2013.09.008

White, J. (2017). Accounting for the learnability of saltation in phonological theory: a maximum entropy model with a P-map bias. *Language* 93, 1–36. doi: 10.1353/lan.2017.0001

Wilson, C. (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cogn. Sci*. 30, 945–982. doi: 10.1207/s15516709cog0000_89

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc*. 73, 3–36. doi: 10.1111/j.1467-9868.2010.00749.x

Yildiz, Y. (2005). "The structure of initial /s/-clusters: evidence from L1 and L2 acquisition," in *Developmental Paths in Phonological Acquisition*, eds M. Tzakosta, C. Levelt, and J. van der Weijer (Leiden: LUCL), 163–187.

Young, E. D. (2008). Neural representation of spectral and temporal information in speech. *Philos. Trans. R. Soc. B: Biol. Sci*. 363, 923–945. doi: 10.1098/rstb.2007.2151

Zuidema, W., and de Boer, B. (2009). The evolution of combinatorial phonology. *J. Phonet*. 37, 125–144. doi: 10.1016/j.wocn.2008.10.003

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.