Measuring the Perceptual Availability of Phonological Features During Language Acquisition Using Unsupervised Binary Stochastic Autoencoders

Cory Shain
Department of Linguistics
The Ohio State University
shain.3@osu.edu

Micha Elsner
Department of Linguistics
The Ohio State University
melsner@ling.osu.edu

Abstract

In this paper, we deploy binary stochastic neural autoencoder networks as models of infant language learning in two typologically unrelated languages (Xitsonga and English). We show that the drive to model auditory percepts leads to latent clusters that partially align with theory-driven phonemic categories. We further evaluate the degree to which theory-driven phonological features are encoded in the latent bit patterns, finding that some (e.g. [±approximant]), are well represented by the network in both languages, while others (e.g. [±spread glottis]) are less so. Together, these findings suggest that many reliable cues to phonemic structure are immediately available to infants from bottom-up perceptual characteristics alone, but that these cues must eventually be supplemented by top-down lexical and phonotactic information to achieve adult-like phone discrimination. Our results also suggest differences in degree of perceptual availability between features, yielding testable predictions as to which features might depend more or less heavily on top-down cues during child language acquisition.

1 Introduction

Distinctive features like [±voice] and [±sonorant] have been a core construct of phonological theory for many decades (Trubetskoy, 1939; Jakobson et al., 1951; Chomsky and Halle, 1968; Clements, 1985). They have been used in automatic speech recognition (Livescu and Glass, 2004), and psycholinguistic evidence suggests that they are cognitively available during language acquisition (Kuhl, 1980; White and Morgan, 2008). Nonetheless, distinctive features are not directly observed by humans; they are abstractions that must be inferred from dense perceptual information (sound waves) during language acquisition and comprehension, which raises questions about how they are learned and recognized. In adults, phonological comprehension is aided by top-down lexical and phonotactic (i.e. sound sequencing) constraints. For example, the classic *phonemic restoration effect* (Warren, 1970) shows that adults infer missing phonemes from context with such ease that they often fail to notice when acoustic cues to phone identity are erased. However, infants first learning their phonemic categories have not yet acquired reliable top-down lexical and phonotactic models and must rely more heavily on bottom-up perceptual information. To a learner faced with the immense challenge of discovering structure in dense perceptual input, do theory-driven phonological features "stand out" or are they swamped by noise? In this paper, we address this question using an unsupervised computational acquisition model.

Previous models of phonological category induction have emphasized the importance of top-down information (information about the contexts in which phonemes occur) (Peperkamp et al., 2006; Swingley, 2009; Feldman et al., 2009a, 2013a,b; Moreton and Pater, 2012a,b; Martin et al., 2013; Pater and Moreton, 2014; Frank et al., 2014; Doyle et al., 2014; Doyle and Levy, 2016). But to prevent the acquisition process from being circular, the learner cannot operate solely on top-down information — the acoustic signal must provide some evidence for the phonemic categories. We hypothesize that the same must be true for at least some phonological features (e.g. [±nasal], [±lateral]), but previous work on unsupervised speech processing has inferred phonological structure from spoken utterances using either (1) discrete transition-based architectures (Varadarajan et al., 2008; Jansen and Church, 2011; Lee and Glass, 2012), which do attempt to discover featurally-related natural classes, or (2) continuous deep neural (Kamper et al., 2015,

Proceedings of NAACL-HLT 2019, pages 69–85
Minneapolis, Minnesota, June 2 - June 7, 2019.

2017a; Renshaw et al., 2015) architectures, whose internal representations are difficult to interpret. Furthermore, these approaches do not separate the contributions of top-down sequential information from bottom-up acoustic properties of segments, making it difficult to assess the relative importance of these information sources throughout the acquisition process.

By contrast, our model attends exclusively to phone-internal acoustic patterns using a deep neural autoencoder with a discrete embedding space composed of binary stochastic neurons (BSNs) (Rosenblatt, 1958; Hinton, 2012; Bengio et al., 2013; Courbariaux et al., 2016). BSNs allow us to exploit (1) the interpretability of discrete representations, (2) the decomposability of phone segments into phonological features, and (3) and the power of deep neural function approximators to relate percepts and their representations. Since every token is labeled with a binary latent code, it is possible to evaluate the model's recovery not only of phonological categories but also of phonological features. Featural representations can encode distributional facts about which processes apply to which classes of sounds in ways that cross-cut the phonological space, rather than simply grouping each segment with a set of similar neighbors (LeCun et al., 2015). By focusing on the acoustic properties of sounds themselves rather than their sequencing in context, our model enables exploration of two questions about the data available to young learners whose training signal must primarily be extracted from bottom-up perceptual information: (1) to what extent can phoneme categories emerge from a drive to model auditory percepts, and (2) how perceptually available are theory-driven phonological features (that is, how easily can they be extracted directly from low-level acoustic percepts)?

Our results show (a) that phonemic categories emerge naturally but imperfectly from perceptual reconstruction and (b) that theory-driven features differ in their degree of perceptual availability. Together, these findings suggest that many reliable cues to phonemic structure are immediately available to infants from bottom-up perceptual characteristics alone, but that these cues may eventually need to be supplemented by top-down lexical and phonotactic information to achieve adult-like phone discrimination (Feldman et al., 2013a; Pater and Moreton, 2014). Our findings also suggest hypotheses as to precisely which kinds of phonological features are more or less perceptually available and therefore might depend more or less heavily on top-down cues for acquisition. Such differences might suggest relative timelines at which different features might be appropriated in support of phonemic, phonotactic, and lexical generalization, providing a rich set of testable hypotheses about child language acquisition.

2 Background and Related Work

2.1 Unsupervised Speech Processing

The present paper has a strong connection to recent work on unsupervised speech processing, especially the Zerospeech 2015 (Versteegh et al., 2015) and 2017 (Dunbar et al., 2017) shared tasks. Participating systems (Badino et al., 2015; Renshaw et al., 2015; Agenbag and Niesler, 2015; Chen et al., 2015; Baljekar et al., 2015; Räsänen et al., 2015; Lyzinski et al., 2015; Zeghidour et al., 2016; Heck et al., 2016; Srivastava and Shrivastava, 2016; Kamper et al., 2017b; Chen et al., 2017; Yuan et al., 2017; Heck et al., 2017; Shibata et al., 2017; Ansari et al., 2017a,b) perform unsupervised ABX discrimination and/or spoken term discovery on the basis of unlabeled speech alone. The design and evaluation of these and related systems (Kamper et al., 2015, 2017a; Elsner and Shain, 2017; Räsänen et al., 2018) are oriented toward word-level modeling. As such, our focus on the perceptual availability of phonological features is orthogonal to but complementary with this line of research. Since distinctive features are important for indexing lexical contrasts, especially between highly confusable words (e.g. onset voicing alone distinguishes *sap* and *zap* in English), studying the perceptual availability of distinctive features to an unsupervised learner may help improve the design and analysis of low-resource speech processing systems.

To our knowledge, the task most closely related to the current paper is unsupervised phone discovery. Some studies in this tradition segment speech into phone-like units without clustering them (Dusan and Rabiner, 2006; Qiao et al., 2008), while others cluster small subsets of pre-segmented sounds (usually vowels) using parametric models (mixture-of-Gaussians) (Vallabha et al., 2007; Feldman et al., 2013a; Antetomaso et al., 2017). Further work combines these tasks and extends the approach to cover the entire acoustic space (Lee and Glass, 2012). However, for a variety of reasons, the Lee and Glass (2012) model does not straightforwardly support evaluation of the perceptual availability of phonological features. First, they do not quantitatively evaluate the discovered phoneme clusters. Second, the model incorporates phonotactics through transition probabilities, making it difficult to disentangle the contributions of top-down and bottom-up information to the learning process. Third, the clustering model is not feature-based, but instead consists of atomic categories, each defining a distinct generative process for acoustics. This design is at odds with the widely held view in linguistic theory that phonemes are not inscrutable atoms of the phonological grammar, but instead labels for bundles of features that define natural classes (Clements, 1985). Our approach is therefore more appropriate to the question at hand.

2.2 Distinctive Features and Phonology Acquisition

There is a great deal of evidence that many phonological contrasts are perceptually available from a very early stage (Eimas et al., 1971; Moffitt, 1971; Trehub, 1973; Jusczyk and Derrah, 1987; Eimas et al., 1987). However, studies of infant phone discrimination typically use carefully-enunciated laboratory stimuli, which have been shown to be substantially easier to discriminate than phones in naturalistic utterances (Feldman et al., 2013a; Antetomaso et al., 2017). It is thus likely that inferring phone categories from acoustic evidence is a persistently challenging task, and studies have found language-specific tuning of the speech perception system from fetal stages (Moon et al., 2013) through the first year (Kuhl et al., 1992; Werker and Tees, 1984) and even all the way into the preteen years (Hazan and Barrett, 2000).

Experiments show that these contrasts are expressed, not simply as oppositions between particular categories, but as a featural system, even in early infancy. Evidence of featural effects has been found in the phone discrimination patterns of both adults (Chládková et al., 2015) and infants (Kuhl, 1980; Hillenbrand, 1985; White and Morgan, 2008). Studies have also shown that infants generalize new distinctions along featural dimensions (Maye et al., 2008b; Cristià et al., 2011). Given infants' early detection and use of some featural contrasts, we hypothesize that there is strong evidence in the acoustic signal for these distinctions, which may then bootstrap the acquisition of phonotactic and lexical patterns (Beckman and Edwards, 2000).

Experiments also suggest asymmetries in the perceptual availability of features. For example, a consonant-vowel distinction appears to be an important early foothold in phonology acquisition: vowel/consonant discrimination emerges early in infant speech processing (Dehaene-Lambertz and Dehaene, 1994), language-specificity in perception follows different timecourses for consonants (Werker and Tees, 1984) and vowels (Kuhl et al., 1992), and vowels and consonants play distinct roles in lexical access vs. rule discovery in children (Nazzi, 2005; Pons and Toro, 2010; Hochmann et al., 2011). Young infants have also been shown to be sensitive to voicing contrasts (Lasky et al., 1975; Aslin et al., 1981; Maye et al., 2008b). Features that distinguish consonant-like from vowel-like segments or voiced from unvoiced segments may thus be highly available to young learners. Infants struggle by comparison with other kinds of phone discrimination tasks, including certain stop-fricative contrasts (Polka et al., 2001) and certain place distinctions within nasal (Narayan et al., 2010) and sibilant (Nittrouer, 2001; Cristià et al., 2011) segments. Even adults struggle with fricative place discrimination from strictly acoustic cues (McGuire and Babel, 2012). Similar asymmetries emerge from our unsupervised learner, as shown in Section 4.2.

Our computational acquisition model complements this experimental research in several ways. First, its internal representations, unlike those of human infants, are open to detailed analysis, even when exposed to naturalistic language stimuli. Second, we can perform cross-linguistic comparisons using readily available corpora without requiring access to a pool of human subjects in each language community. Third, our model provides global and graded quantification of the perceptual availability of distinctive features in natural speech, permitting us to explore relationships between features in a way that is difficult to do through experiments on infants, which are generally constrained to same-different contrasts over a small set of manipulations.

2.3 Cognition and the BSN Autoencoder

The reconstruction objective used here is not merely a convenient supervision signal. There is reason to believe that people actively model their perceptual worlds (Mamassian et al., 2002; Feldman, 2012; Singer et al., 2018; Yan et al., 2018), and autoassociative structures have been found in several brain areas (Treves and Rolls, 1991; Rolls and Treves, 1998). There is also evidence that phonetic comprehension and production can be acquired symbiotically through a sensorimotor loop relating acoustic perception and articulator movements (Houde and Jordan, 1998; Fadiga et al., 2002; Watkins et al., 2003; Wilson et al., 2004; Pulvermüller et al., 2006; Kröger et al., 2009; Bolhuis et al., 2010; Kröger and Cao, 2015; Bekolay, 2016). Finally, evidence suggests that working memory limitations impose compression pressures on the perceptual system that favor sparse representations of dense acoustic percepts (Baddeley and Hitch, 1974) and may guide infant language acquisition (Baddeley et al., 1998; Elsner and Shain, 2017). It is thus reasonable to suppose that perceptual reconstruction — such as that implemented by an autoencoder architecture — is immediately available as a learning signal to infants who still lack reliable guidance from phonotactics or the lexicon.

Our use of BSNs follows the spirit of the earliest work on artificial neural networks (Rosenblatt, 1958). Rosenblatt's perceptron was designed to study learning and decision-making in the brain and therefore used binary neurons to model the discrete firing behavior of their biological counterparts. This tradition has been replaced in deep learning research with differentiable activation functions that support supervised learning through backpropagation of error but are less biologically plausible. Our work takes advantage of the development of effective estimators for the gradients of discrete neurons (Williams, 1992; Hinton, 2012; Bengio et al., 2013; Courbariaux et al., 2016; Chung et al., 2017) to wed these two traditions, exploiting BSNs to encode the learner's latent representation of auditory percepts and deep networks to map between percepts and their latent representations. In addition to the greater similarity of BSNs to biological neurons, the use of discrete featural representations is motivated by experimental evidence that human phone perception (including that of infants) is both featural (White and Morgan, 2008; Chládková et al., 2015) and categorical (Liberman et al., 1961; Eimas et al., 1987; Harnad, 2003; Feldman et al., 2009b).

Experiments reported here use an 8-bit binary segment encoding. Eight bits is the the lower bound on binary encodings that are sufficiently expressive to capture all segmental contrasts in any known language (Mielke, 2009). Although theory-driven taxonomies generally contain more than eight distinctive features, these taxonomies are known to be highly redundant (Cherry et al., 1953). For example, the phonological featurization of the Xitsonga segments analyzed in our experiments contains 26 theory-driven features (Hayes, 2011; Hall et al., 2016), yielding up to $2^{26} = 67108864$ distinct segment categories, far more than the number of known segment types in Xitsonga or even the number of training instances in our data. By entailment, any representation that can identify all segment types in a language can also identify all featural contrasts that discriminate those types, regardless of how the feature space is factored. For this reason, we consider a phonological feature to be represented if it can be detected by an arbitrary function of the latent bits (Section 4.2), without assuming that the true and discovered feature spaces will factor identically.

2.4 Supervised Acoustic Feature Learning

Our study shares an interest in phonological features with previous work in automatic speech recognition attempting to discover mappings between acoustics and hand-labeled featural representations (Liu, 1996; Bitar and Espy-Wilson, 1996; Frankel and King, 2001; Kirchhoff et al., 2002; Livescu and Glass, 2004; Mitra et al., 2011, *inter alia*). While these results provide evidence that such a mapping is indeed learnable in an oracle setting, they rely on a supervision signal (direct annotation of the target representations) to which children do not have access. Our unsupervised approach measures perceptual availability of features in a more realistic learning scenario.

3 Experimental Setup

3.1 Model

The simulated learner used in this study is a deep neural autoencoder with an 8-bit layer of BSNs as its principle information bottleneck, depicted in Figure 1. The model processes a given phone segment by encoding the segment's acoustic informa-

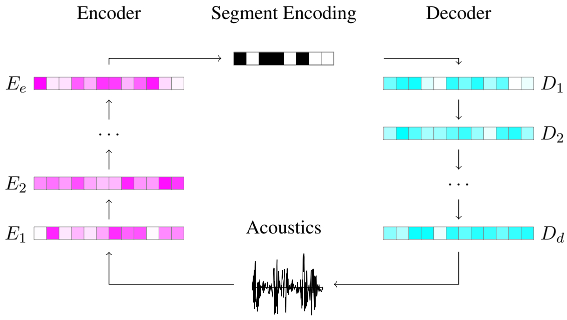


Figure 1: The binary stochastic neural autoencoder architecture with encoder layers $E_{1,...,e}$ and decoder layers $D_{1,...,d}$. For expository purposes, acoustics are represented as pressure waves. In reality, the system uses frames of Mel frequency cepstral coefficients.

tion into a bit pattern and then reconstructing the acoustic information from the encoded bit pattern. It is thus incentivized to use the latent bits in a systematic featural manner, encoding similar segments in similar ways.

The encoder and decoder are both deep feedfoward residual networks (He et al., 2016).[1] To enable feedforward autoencoding of sequential data, phone segments are clipped at 50 timesteps (500ms), providing complete coverage of over 99% of the phone segments in each corpus. Given F-dimensional input acoustic frames and a maximum input length of M timesteps, the weight matrix of each encoder layer is $\in \mathbb{R}^{FM \times FM}$ except the final layer ($\in \mathbb{R}^{FM \times 8}$). Given R-dimensional reconstructed acoustic frames and a maximum output length of N timesteps, the weight matrix of each decoder layer is $\in \mathbb{R}^{RN \times RN}$ except the first layer ($\in \mathbb{R}^{8 \times RN}$). Both the encoder and decoder contain initial and final dense transformation layers, with three residual layers in between. Each residual layer contains two dense layers. All internal layers use tanh activations and are batch-normalized with a decay rate of 0.9 (Ioffe and Szegedy, 2015).

Given that the capacity for *speaker adaptation* — short-term accommodation of idiosyncrasies in individuals' productions — has been shown for both adults (Clarke and Garrett, 2004; Maye et al., 2008a) and children (Kuhl, 1979; van Heugten and Johnson, 2014), we equip the models with a 16-dimensional speaker embedding, which is concatenated both to the acoustic input frames and to the latent bit vector.

Each BSN of the latent encoding is associated with a firing probability $\in [0, 1]$ parameterized by the encoder network. The neural activation can be discretized either deterministically or by sampling. The use of BSNs to encode segments is a problem for gradient-based optimization since it introduces a non-differentiable discrete decision into the network's latent structure. We overcome this problem by approximating missing gradients using the straight-through estimator (Hinton, 2012; Bengio et al., 2013; Courbariaux et al., 2016) with slope-annealing (Chung et al., 2017). Slope annealing multiplies the pre-activations a by a monotonically increasing function of the training iteration t, incrementally decreasing the bias of the straight-through estimator. We use the following annealing function:

$$a \leftarrow a(1 + 0.1t)$$

We discretize the latent dimensions using Bernoulli sampling during training and thresholding at 0.5 during evaluation.

The models are implemented in Tensorflow (Abadi et al., 2015) and optimized using Adam (Kingma and Ba, 2014) for 150 training epochs with a constant learning rate of 0.001. The source code is available at `https://github.com/coryshain/dnnseg`.

3.2 Data

We apply our model to the Xitsonga and English speech data from the Zerospeech 2015 shared task. The Xitsonga data are drawn from the NCHLT corpus (De Vries et al., 2014) and contain 2h29m07s of read speech from 24 speakers. The English data are drawn from the Buckeye Corpus (Pitt et al., 2005) and contain 4h59m05s of conversational speech from 12 speakers. While neither of these corpora represent child-directed speech, they both consist of fluently produced word tokens in context, rather than isolated productions as in many previous laboratory studies with infants (Eimas et al., 1971; Werker and Tees, 1984; Kuhl et al., 1992, *inter alia*). We pre-segment the audio files using time-aligned phone transcriptions pro-

[1] Feedforward networks are used both for computational reasons and because they dramatically outperformed recurrent networks in initial experiments, especially when RNN's were used for decoding. We hypothesize that this is due to the lack of direct access to the encoder timesteps, such as that permitted by sequence to sequence models with attention (Bahdanau et al., 2015). Attention is not viable for our goals because it defeats the purposes of an autoencoder by allowing the decoder to bypass the encoder's latent representation.

Model	Xitsonga H	Xitsonga C	Xitsonga V	English H	English C	English V
Baseline	0.023	0.013	0.016	0.006	0.004	0.005
Sigmoid	0.281	0.191	0.227	0.246	0.166	0.198
Sigmoid+Speaker	0.302	0.185	0.230	0.205	0.180	0.192
BSN	0.360	0.206	0.262	0.240	0.161	0.193
Our model (BSN+Speaker)	**0.462**	**0.268**	**0.339**	**0.270**	**0.180**	**0.216**

Table 1: Phone clustering scores. Homogeneity (H), completeness (C) and V-measure (V) across the Zerospeech 2015 Xitsonga and English challenge datasets.

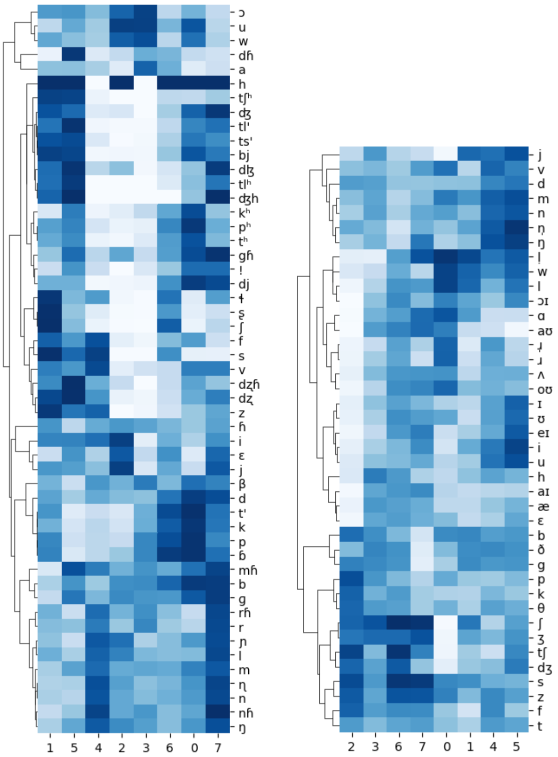
vided in the challenge repository. The gold segment labels are used in clustering evaluation metrics, but the unsupervised learner never has access to them. Data selection criteria and annotation procedures are are described in more detail in Versteegh et al. (2015).

Prior to fitting, we apply a standard spectral preprocessing pipeline from automatic speech recognition: raw acoustic signals are converted into 13-dimensional vectors of Mel frequency cepstral coefficients (MFCCs) (Mermelstein, 1976) with first and second order deltas, yielding 39-dimensional frames sequenced in time. Each frame covers 25ms of speech, and frames are spaced 10ms apart. The deltas are used by the encoder but stripped from the reconstruction targets. Following preceding work showing improved unsupervised clustering when segments are given fixed-dimensional acoustic representations, thus abstracting away from the variable temporal dilation in natural speech (Kamper et al., 2017a,b), we resample all reconstruction targets to a length of 25 frames.

This pipeline instantiates some standard assumptions about the perceptual representations underlying human speech processing. Alternative representations — for instance, articulatory representations (Liu, 1996; Frankel and King, 2001; Kirchhoff et al., 2002; Livescu and Glass, 2004) or other spectral transforms (Zwicker, 1961; Makhoul, 1975; Hermansky, 1990; Hermansky et al., 1991; Coifman and Wickerhauser, 1992; Shao et al., 2009) — have been proposed as alternatives to MFCCs. Our results concerning perceptual availability are of course tied to our input representation, since phenomena that are poorly distinguished by MFCCs have less effect on our autoencoder loss function. Nonetheless, MFCCs are known to produce high-quality supervised speech recognizers (Zheng et al., 2001; Hinton et al., 2012), and we therefore leave optimization of the representation of speech features to future work.

4 Results and Discussion

4.1 Phonemic Categories Partially Emerge from Modeling Auditory Percepts



(a) Xitsonga

(b) English

Figure 2: Mean activation pattern by gold segment label from the BSN model with speaker embeddings, with darker color indexing higher average activation.

The first research question posed in the introduction was to what extent theory-driven phoneme categories emerge from a drive to model auditory percepts. We explore this question by evaluating the degree of correspondence between the autoencoder hidden states and the gold phone labels. Table 1 reports learning outcomes using the information theoretic measures *homogeneity* (H), *completeness* (C), and *V-measure* (V) for unsupervised cluster evaluation (Rosenberg and Hirschberg, 2007). All three metrics range over the interval $[0, 1]$, with 1 indexing perfect performance. As shown in the table, our model yields dramatically better clustering performance than a random baseline that uniformly draws cluster IDs from a pool of 256 categories: we obtain 2118% and 4500% relative V-measure improvements in Xitsonga and English, respectively. At the same time, clustering performance is far from perfect. This result indicates that perceptual modeling — an immediately-available learning signal in infant language acquisition — both (1) drives the learner a long way toward phoneme acquisition, and (2) is insufficient to fully identify phone categories in our learners. One likely explanation for the latter is evidence from cognitive science that phonotactic and lexical information (to which our learners do not have access) supplement perception as the acquisition process unfolds (Feldman et al., 2013a; Pater and Moreton, 2014).

The middle rows of Table 1 show ablation results from using non-discrete sigmoid neurons rather than BSNs in the encoding layer (*Sigmoid* vs. *BSN*)[2] and/or removing the speaker adaptation feature (i.e. removing speaker embeddings). As shown, the classification performance of our model benefits substantially from the use of BSN encodings with speaker adaptation, especially on Xitsonga. Note that the reconstruction losses of the sigmoid encoders are better than those of the BSN encoders despite their degraded classification performance. This is to be expected: sigmoid neurons have greater representational capacity than binary neurons, since they can encode information through continuous gradations. They are therefore more capable of memorizing idiosyncratic properties of the input and are less incentivized to discover generalizable latent classes. The ablation results thus suggest that speaker adaptation and categorical perception support the discovery of linguistically relevant abstractions.

4.2 Distinctive Features Differ in Perceptual Availability

The second research question posed in the introduction was to what extent distinctive features differ in perceptual availability. We explore this question in two ways.

First, we qualitatively assess the linguistic plausibility of the natural clustering in the latent bits. Figure 2 visualizes this clustering based on correlations between the average of the bit patterns across all instances of each gold phone type for both datasets. If the unsupervised classifier ignored phonological structure altogether, the plots would be roughly uniform in color, and if the unsupervised classifier perfectly identified phonemes, the plots would consist entirely of fully light or fully dark cells, with unique bit patterns associated with each phone type. As shown, the reality falls in between: while the visualized classifications are far from perfect, they nonetheless contain a great deal of structure and suggest the presence of rough natural classes in both languages, especially of affricates, nasals, sibilants, and approximants. Our learners also replicate infants' difficulty in discriminating some nasal and fricative place features (Polka et al., 2001; Nittrouer, 2001; Narayan et al., 2010), assigning highly similar representations to many subtypes of nasals and fricatives across places of articulation (see e.g. similar mean bit patterns of /n/ vs. /ŋ/ and /s/ vs. /ʃ/ in both languages).

Second, we quantitatively evaluate the degree to which theory-driven features like [±voice] are recoverable from the network's latent representations. To do so, we map gold phone labels into binary distinctive feature clusters from Hayes (2011) using Phonological CorpusTools (Hall et al., 2016). One possible form of analysis would be to search for individual correspondences between distinctive features and the model's latent dimensions. However, this is likely to underestimate the degree of feature learning because the deep decoder can learn arbitrary logics on the latent bit patterns, a necessary property for fitting complex non-linear mappings from latent features to acoustics. We instead evaluate distinctive feature discovery by fitting random forest classifiers that predict theory-driven features using the latent bit patterns as inputs. We can then use classifier performance to assess the degree to which a given distinctive feature can be recovered by a logical statement on the network's latent bits. The classifiers were fitted using 5-fold cross-validation in Scikit-learn (Pedregosa et al., 2011) with 100 estimators, balanced class weighting, and an entropy-based split criterion.

Results are given in Tables 2 and 3. As shown, (1) there are large differences in perceptual availability between features, and (2) relative avail-

[2] To obtain class labels from the sigmoid encoder, we rounded the activations. Rounding was only used for evaluation and had no impact on the fitting procedure.

Feature	P	R	F
voice	0.9767	0.9033	0.9386
sonorant	0.9249	0.9085	0.9166
continuant	0.9492	0.7936	0.8645
consonantal	0.8314	0.8915	0.8604
approximant	0.8998	0.8192	0.8576
syllabic	0.8278	0.8523	0.8398
dorsal	0.8935	0.7703	0.8273
strident	0.6991	0.9594	0.8089
low	0.7175	0.8978	0.7976
front	0.6590	0.8101	0.7268
high	0.5875	0.7882	0.6732
back	0.5352	0.8527	0.6577
round	0.5332	0.8551	0.6568
labial	0.5669	0.7725	0.6539
coronal	0.5382	0.8301	0.6530
tense	0.5208	0.8115	0.6344
delayed release	0.5468	0.7226	0.6225
anterior	0.4078	0.8355	0.5481
nasal	0.3635	0.8796	0.5144
distributed	0.2459	0.8537	0.3819
constricted glottis	0.1762	0.9007	0.2948
lateral	0.1536	0.8062	0.2581
labiodental	0.0934	0.7980	0.1672
trill	0.0809	0.7401	0.1458
spread glottis	0.0671	0.5856	0.1204
implosive	0.0041	0.4041	0.0081

Table 2: Perceptual availability by feature in Xitsonga

Feature	P	R	F
voice	0.9244	0.8567	0.8893
sonorant	0.8544	0.8862	0.8700
approximant	0.8005	0.8370	0.8183
continuant	0.8577	0.7669	0.8098
consonantal	0.8249	0.7357	0.7777
syllabic	0.6624	0.8426	0.7417
dorsal	0.7046	0.7114	0.7080
strident	0.5505	0.9027	0.6839
coronal	0.5758	0.7066	0.6345
anterior	0.5251	0.7280	0.6101
delayed release	0.4413	0.7374	0.5521
front	0.4322	0.7407	0.5459
high	0.3841	0.6931	0.4943
tense	0.3275	0.7101	0.4483
back	0.3128	0.7504	0.4416
nasal	0.2796	0.7544	0.4080
labial	0.2541	0.7077	0.3739
low	0.2410	0.7787	0.3680
distributed	0.2203	0.6881	0.3337
diphthong	0.2039	0.8051	0.3254
round	0.1665	0.7012	0.2692
lateral	0.1484	0.8333	0.2519
labiodental	0.0787	0.6756	0.1410
spread glottis	0.0377	0.6683	0.0714

Table 3: Perceptual availability by feature in English

ability of features is remarkably consistent between these unrelated languages, suggesting that the models are tapping into generalized perceptual patterns. The best-learned feature in both languages is [±voice], which is consistent with early evidence of voicing sensitivity in infants (see Section 2.2). Below this, the features [±sonorant], [±continuant], [±consonantal], [±approximant], and [±syllabic] are faithfully recovered in both languages. All of these features distinguish prototypical consonants from prototypical vowels but differ in their treatment of edge cases like nasals, liquids, and glides. Thus, similarly to the infant subjects discussed in Section 2.2, the model finds the consonant-vowel contrast to be highly available. Like human infants, our computational learner finds certain consonantal place and manner features relatively more difficult, although the features [±dorsal], [±coronal], [±strident] and [±delayed release] are also fairly well recovered in both languages. By contrast, both models poorly capture features like [±lateral], [±labiodental], [±distributed], [±nasal], [±constricted glottis], [±spread glottis], and [±implosive],[3] suggesting that these features are more difficult to discover bottom-up and may therefore be more dependent on phonotactic and lexical constraints for acquisition.[4] This finding aligns with the acquisition literature in suggesting that there may be substantial differences in perceptual availability between different place and manner features (see Section 2.2).

In addition to these cross-linguistic similarities, the models also reveal important differences between Xitsonga and English. For example, the two languages differ in the relative availability of features that distinguish vowels vs. features that distinguish consonants. In English, vowel features like [±front], [±high], and [±back] are substantially less well learned than consonant features like [±coronal], [±anterior], and [±delayed release], while the opposite holds in Xitsonga. We hypothesize that this is due to the fact that there are more vowels and fewer consonants in English than in Xitsonga: having fewer distinctions might reduce the degree of “crowding“ in the articulatory space, increasing perceptual contrast between phone types (Liljencrants and Lindblom, 1972).

[3] Delayed release: affricates, constricted glottis: ejectives; spread glottis: glottal frication (e.g. aspirated stops).

[4] Note that we are not suggesting that e.g. [±spread glottis] cannot be detected in speech. Our claim is rather that acoustic cues to [±spread glottis] are less pronounced and/or less reliable than cues to e.g. [±voice] and therefore perhaps more difficult to exploit in early infancy, since our autoencoder model does not find them particularly useful for perceptual reconstruction.

Finally, note that the cluster maps in Figure 2 and the feature recovery data in Tables 2 and 3 provide complementary perspectives on the learned representations. For example, it may at first seem surprising that the feature [±nasal] is recovered relatively poorly in both languages, given that nasals are well clustered in Figure 2. This discrepancy indicates that nasal segments are represented similarly to each other but also similarly enough to other segments that they are not reliably differentiated as a class. Conversely, the voicing feature is well recovered in both languages despite the lack of a visible cluster of voiced segments. This indicates that voicing is reliably encoded in the latent bits, even if the representation as a whole is dominated by other kinds of information.

5 Conclusion

In this paper, we used binary stochastic neural autoencoders to explore the perceptual availability of (1) theory-driven phonemic categories and (2) theory-driven phonological features, based only on the acoustic properties of segments. We found that phonemic categories exert substantial influence on a learner driven to model its auditory percepts, but that additional information — especially phonotactic and lexical (Feldman et al., 2013a) — is likely necessary for full adult-like phone discrimination. We also found asymmetries in the perceptual availability of phonological features like [±voice] and [±nasal] and showed that these asymmetries reflect attested patterns of infant phone discrimination. Our model both replicates broad trends in the child acquisition literature (successful consonant-vowel and voicing discrimination, relatively less successful discrimination of various place and manner features) and sheds new light on potential relationships between auditory perception and language acquisition: the overall cline of perceptual availability revealed by the model in Tables 2 and 3 suggests a range of testable hypotheses about the role of perception in infant speech processing.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by National Science Foundation grant #1422987 to ME. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Wiehan Agenbag and Thomas Niesler. 2015. Automatic segmentation and clustering of speech using sparse coding and metaheuristic search. In *Sixteenth Annual Conference of the International Speech Communication Association*.

T K Ansari, Rajath Kumar, Sonali Singh, and Sriram Ganapathy. 2017a. Deep learning methods for unsupervised acoustic modelingLeap submission to ZeroSpeech challenge 2017. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 754–761. IEEE.

T K Ansari, Rajath Kumar, Sonali Singh, Sriram Ganapathy, and Susheela Devi. 2017b. Unsupervised HMM posteriograms for language independent acoustic modeling in zero resource conditions. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 762–768. IEEE.

Stephanie Antetomaso, Kouki Miyazawa, Naomi Feldman, Micha Elsner, Kasia Hitczenko, and Reiko Mazuka. 2017. Modeling phonetic category learning from natural acoustic data. In *Proceedings of the annual Boston University Conference on Language Development*.

Richard N Aslin, David B Pisoni, Beth L Hennessy, and Alan J Perey. 1981. Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child development*, 52(4):1135.

Alan Baddeley, Susan Gathercole, and Costanza Papagno. 1998. The Phonological Loop as a Language Learning Device. *Psychological Review*, 105(1):158–173.

Alan D Baddeley and Graham Hitch. 1974. *Working Memory*. University of Stirling, Stirling, Scotland.

Leonardo Badino, Alessio Mereta, and Lorenzo Rosasco. 2015. Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations 2015*.

Pallavi Baljekar, Sunayana Sitaram, Prasanna Kumar Muthukumar, and Alan W Black. 2015. Using articulatory features and inferred phonological segments in zero resource speech processing. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Mary E Beckman and Jan Edwards. 2000. The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child development*, 71(1):240–249.

Trevor Bekolay. 2016. *Biologically inspired methods in speech recognition and synthesis: Closing the loop*. Ph.D. thesis, University of Waterloo.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Nabil N Bitar and Carol Y Espy-Wilson. 1996. Knowledge-based parameters for HMM speech recognition. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 29–32. IEEE.

Johan J Bolhuis, Kazuo Okanoya, and Constance Scharff. 2010. Twitter evolution: converging mechanisms in birdsong and human speech. *Nature Reviews Neuroscience*, 11(11):747.

Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li. 2015. Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li. 2017. Multilingual bottle-neck feature learning from untranscribed speech. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 727–733. IEEE.

E Colin Cherry, Morris Halle, and Roman Jakobson. 1953. Toward the logical description of languages in their phonemic aspect. *Language*, pages 34–46.

Katerina Chládková, Paul Boersma, Titia Benders, and others. 2015. The perceptual basis of the feature vowel height. In *ICPhS*.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper \& Row.

Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. 2017. Hierarchical Multiscale Recurrent Neural Networks. In *International Conference on Learning Representations 2017*.

Constance M Clarke and Merrill F Garrett. 2004. Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6):3647–3658.

George N Clements. 1985. The geometry of phonological features. *Phonology*, 2(1):225–252.

Ronald R Coifman and M Victor Wickerhauser. 1992. Entropy-based algorithms for best basis selection. *IEEE Transactions on information theory*, 38(2):713–718.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*.

Alejandrina Cristià, Grant L McGuire, Amanda Seidl, and Alexander L Francis. 2011. Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of phonetics*, 39(3):388–402.

Nic J De Vries, Marelie H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard, and Alta De Waal. 2014. A smartphone-based ASR data collection tool for under-resourced languages. *Speech communication*, 56:119–131.

Ghislaine Dehaene-Lambertz and Stanislas Dehaene. 1994. Speed and cerebral correlates of syllable discrimination in infants. *Nature*, 370(6487):292.

Gabriel Doyle, Klinton Bicknell, and Roger Levy. 2014. Nonparametric learning of phonological constraints in optimality theory. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1094–1103.

Gabriel Doyle and Roger Levy. 2016. Data-driven learning of symbolic constraints for a log-linear model in a phonological setting. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2217–2226.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 323–330. IEEE.

Sorin Dusan and Lawrence Rabiner. 2006. On the relation between maximum spectral transition positions and phone boundaries. In *Ninth International Conference on Spoken Language Processing*.

Peter D. Eimas, Joanne L. Miller, and Peter W. Jusczyk. 1987. On infant speech perception and the acquisition of language. In Stevan Harnad, editor, *Categorical perception: The groundwork of cognition*, pages 161–195. Cambridge University Press, New York.

Peter D Eimas, Einar R Siqueland, Peter Jusczyk, and James Vigorito. 1971. Speech perception in infants. *Science*, 171(3968):303–306.

Micha Elsner and Cory Shain. 2017. Speech segmentation with a neural encoder model of working memory. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1080.

Luciano Fadiga, Laila Craighero, Giovanni Buccino, and Giacomo Rizzolatti. 2002. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European journal of Neuroscience*, 15(2):399–402.

Jacob Feldman. 2012. Symbolic representation of probabilistic worlds. *Cognition*, 123(1):61–83.

Naomi Feldman, Thomas Griffiths, and James Morgan. 2009a. Learning phonetic categories by learning a lexicon. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Naomi H Feldman, Thomas L Griffiths, Sharon Goldwater, and James L Morgan. 2013a. A role for the developing lexicon in phonetic category acquisition. *Psychological review*, 120(4):751.

Naomi H Feldman, Thomas L Griffiths, and James L Morgan. 2009b. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4):752.

Naomi H Feldman, Emily B Myers, Katherine S White, Thomas L Griffiths, and James L Morgan. 2013b. Word-level information influences phonetic learning in adults and infants. *Cognition*, 127(3):427–438.

Stella Frank, Naomi H Feldman, and Sharon Goldwater. 2014. Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1073–1083.

Joe Frankel and Simon King. 2001. ASR-articulatory speech recognition. In *Seventh European Conference on Speech Communication and Technology*.

Kathleen Currie Hall, Blake Allen, Michael Fry, Scott Mackie, and Michael McAuliffe. 2016. Phonological CorpusTools: A free, open-source tool for phonological analysis. In *14th Conference for Laboratory Phonology*, volume 543.

Stevan Harnad. 2003. Categorical Perception. In *Encyclopedia of Cognitive Science*, volume 67. MacMillan: Nature Publishing Group.

Bruce Hayes. 2011. *Introductory phonology*, volume 32. John Wiley \& Sons, Hoboken.

Valerie Hazan and Sarah Barrett. 2000. The development of phonemic categorization in children aged 6–12. *Journal of phonetics*, 28(4):377–396.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Michael Heck, Sakriani Sakti, and Satoshi Nakamura. 2016. Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario. *Procedia Computer Science*, 81:73–79.

Michael Heck, Sakriani Sakti, and Satoshi Nakamura. 2017. Feature optimized dpgmm clustering for unsupervised subword modeling: A contribution to zerospeech 2017. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 740–746. IEEE.

Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *Second European Conference on Speech Communication and Technology*.

Marieke van Heugten and Elizabeth K Johnson. 2014. Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, 143(1):340.

James Hillenbrand. 1985. Perception of feature similarities by infants. *Journal of Speech & Hearing Research*, 28(2):317–318.

Geoffrey Hinton. 2012. Neural Networks for Machine Learning. *Coursera, video lectures.*

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, and others. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29.

Jean-Rémy Hochmann, Silvia Benavides-Varela, Marina Nespor, and Jacques Mehler. 2011. Consonants and vowels: different roles in early language acquisition. *Developmental science*, 14(6):1445–1458.

John F Houde and Michael I Jordan. 1998. Sensorimotor adaptation in speech production. *Science*, 279(5354):1213–1216.

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, pages 448–456.

Roman Jakobson, C Gunnar Fant, and Morris Halle. 1951. *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT press.

Aren Jansen and Kenneth Church. 2011. Towards unsupervised training of speaker independent acoustic models. In *Twelfth Annual Conference of the International Speech Communication Association*.

Peter W Jusczyk and Carolyn Derrah. 1987. Representation of speech sounds by young infants. *Developmental Psychology*, 23(5):648.

Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. 2015. Unsupervised neural network based feature extraction using weak top-down constraints. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5818–5822. IEEE.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2017a. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174.

Herman Kamper, Karen Livescu, and Sharon Goldwater. 2017b. An embedded segmental k-means model for unsupervised segmentation and clustering of speech. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 719–726. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6.

Katrin Kirchhoff, Gernot A Fink, and Gerhard Sagerer. 2002. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3-4):303–319.

Bernd J Kröger and Mengxue Cao. 2015. The emergence of phonetic–phonological features in a biologically inspired model of speech processing. *Journal of Phonetics*, 53:88–100.

Bernd J Kröger, Jim Kannampuzha, and Christiane Neuschaefer-Rube. 2009. Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793–809.

Patricia K Kuhl. 1979. Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, 66(6):1668–1679.

Patricia K Kuhl. 1980. Perceptual constancy for speech-sound categories in early infancy. *Child phonology*, 2:41–66.

Patricia K Kuhl, Karen A Williams, Francisco Lacerda, Kenneth N Stevens, and Bjrn Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

Robert E Lasky, Ann Syrdal-Lasky, and Robert E Klein. 1975. VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, 20(2):215–225.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.

Chia-ying Lee and James Glass. 2012. A Nonparametric {Bayesian} Approach to Acoustic Model Discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 40–49.

Alvin Liberman, Katherine Safford Harris, Peter Eimas, Leigh Lisker, and Jarvis Bastian. 1961. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 4(4):175–195.

Johan Liljencrants and Bjorn Lindblom. 1972. Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast. *Language*, 48(4):839–862.

Sharlene A Liu. 1996. Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*, 100(5):3417–3430.

Karen Livescu and James Glass. 2004. Feature-based pronunciation modeling for speech recognition. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 81–84. Association for Computational Linguistics.

Vince Lyzinski, Gregory Sell, and Aren Jansen. 2015. An evaluation of graph clustering methods for unsupervised term discovery. In *Sixteenth Annual Conference of the International Speech Communication Association*.

John Makhoul. 1975. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

Pascal Mamassian, Michael Landy, and Laurence T Maloney. 2002. Bayesian modelling of visual perception. In Rajesh P N Rao, Bruno A Olshausen, Michael S Lewicki, Michael I Jordan, and Thomas G Dietterich, editors, *Probabilistic models of the brain: Perception and neural function*, pages 13–36. The MIT Press, Cambridge, MA.

Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. 2013. Learning phonemes with a protolexicon. *Cognitive Science*, 37(1):103–124.

Jessica Maye, Richard N Aslin, and Michael K Tanenhaus. 2008a. The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, 32(3):543–562.

Jessica Maye, Daniel J Weiss, and Richard N Aslin. 2008b. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science*, 11(1):122–134.

Grant McGuire and Molly Babel. 2012. A cross-modal account for synchronic and diachronic patterns of/f/and/$$\theta$$/in English. *Laboratory Phonology*, 3(2):251–272.

Paul Mermelstein. 1976. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388.

Jeff Mielke. 2009. Segment inventories. *Language and linguistics compass*, 3(2):700–718.

Vikramjit Mitra, Hosung Nam, and Carol Espy-Wilson. 2011. Robust speech recognition using articulatory gestures in a Dynamic Bayesian Network framework. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*.

Alan R Moffitt. 1971. Consonant cue perception by twenty-to twenty-four-week-old infants. *Child development*, pages 717–731.

Christine Moon, Hugo Lagercrantz, and Patricia K Kuhl. 2013. Language experienced in utero affects vowel perception after birth: A two-country study. *Acta Paediatrica*, 102(2):156–160.

Elliott Moreton and Joe Pater. 2012a. Structure and substance in artificial-phonology learning, part I: Structure. *Language and linguistics compass*, 6(11):686–701.

Elliott Moreton and Joe Pater. 2012b. Structure and substance in artificial-phonology learning, part II: Substance. *Language and linguistics compass*, 6(11):702–718.

Chandan R Narayan, Janet F Werker, and Patrice Speeter Beddor. 2010. The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*, 13(3):407–420.

Thierry Nazzi. 2005. Use of phonetic specificity during the acquisition of new words: Differences between consonants and vowels. *Cognition*, 98(1):13–30.

Susan Nittrouer. 2001. Challenging the notion of innate phonetic boundaries. *The Journal of the Acoustical Society of America*, 110(3):1598–1605.

Joe Pater and Elliott Moreton. 2014. Structurally biased phonology: complexity in learning and typology. *The EFL Journal*, 3(2).

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Sharon Peperkamp, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Linda Polka, Connie Colantonio, and Megha Sundara. 2001. A cross-language comparison of/d/–/{\dh}/perception: evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, 109(5):2190–2201.

Ferran Pons and Juan M Toro. 2010. Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition*, 116(3):361–367.

Friedemann Pulvermüller, Martina Huss, Ferath Kherif, Fermin Moscoso del Prado Martin, Olaf Hauk, and Yury Shtyrov. 2006. Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, 103(20):7865–7870.

Yu Qiao, Naoya Shimomura, and Nobuaki Minematsu. 2008. Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3989–3992. IEEE.

Okko Räsänen, Gabriel Doyle, and Michael C Frank. 2015. Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Okko Johannes Räsänen, Gabriel Doyle, and Michael C Frank. 2018. Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171:130–150.

Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater. 2015. A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Edmund T Rolls and Alessandro Treves. 1998. *Neural networks and brain function*, volume 572. Oxford University Press, Oxford.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Yang Shao, Zhaozhang Jin, DeLiang Wang, and Soundararajan Srinivasan. 2009. An auditory-based feature for robust speech recognition. In *2009 IEEE*

International Conference on Acoustics, Speech and Signal Processing, pages 4625–4628. IEEE.

Hayato Shibata, Taku Kato, Takahiro Shinozaki, and Shinji Watanabet. 2017. Composite embedding systems for ZeroSpeech2017 Track1. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 747–753. IEEE.

Yosef Singer, Yayoi Teramoto, Ben D B Willmore, Jan W H Schnupp, Andrew J King, and Nicol S Harper. 2018. Sensory cortex is optimized for prediction of future input. *eLife*, 7:e31557.

Brij Mohan Lal Srivastava and Manish Shrivastava. 2016. Articulatory gesture rich representation learning of phonological units in low resource settings. In *International Conference on Statistical Language and Speech Processing*, pages 80–95. Springer.

Daniel Swingley. 2009. Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1536):3617–3632.

Sandra E Trehub. 1973. Infants' sensitivity to vowel and tonal contrasts. *Developmental Psychology*, 9(1):91.

Alessandro Treves and Edmund T Rolls. 1991. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397.

Nikolaï Sergeyevich Trubetskoy. 1939. Grundzüge der phonologie. In *Travaux du Cercle Linguistique de Prague*, volume 7. Van den Hoeck & Ruprecht.

Gautam K Vallabha, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 165–168. Association for Computational Linguistics.

Maarten Versteegh, Roland Thiollière, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The zero resource speech challenge 2015. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Richard M Warren. 1970. Perceptual restoration of missing speech sounds. *Science*, 167(3917):392–393.

Kate E Watkins, Antonio P Strafella, and Tomáš Paus. 2003. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8):989–994.

Janet F Werker and Richard C Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63.

Katherine S White and James L Morgan. 2008. Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, 59(1):114–132.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Stephen M Wilson, Ayşe Pinar Saygin, Martin I Sereno, and Marco Iacoboni. 2004. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7(7):701.

Shaorong Yan, Francis Mollica, and Michael K Tanenhaus. 2018. A context constructivist account of contextual diversity. In *Proceedings of the 40th Annual Cognitive Science Society Meeting*.

Yougen Yuan, Cheung-Chi Leung, Lei Xie, Hongjie Chen, Bin Ma, and Haizhou Li. 2017. Extracting bottleneck features and word-like pairs from untranscribed speech for feature representation. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 734–739. IEEE.

Neil Zeghidour, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. 2016. A deep scattering spectrumdeep siamese network pipeline for unsupervised acoustic modeling. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4965–4969. IEEE.

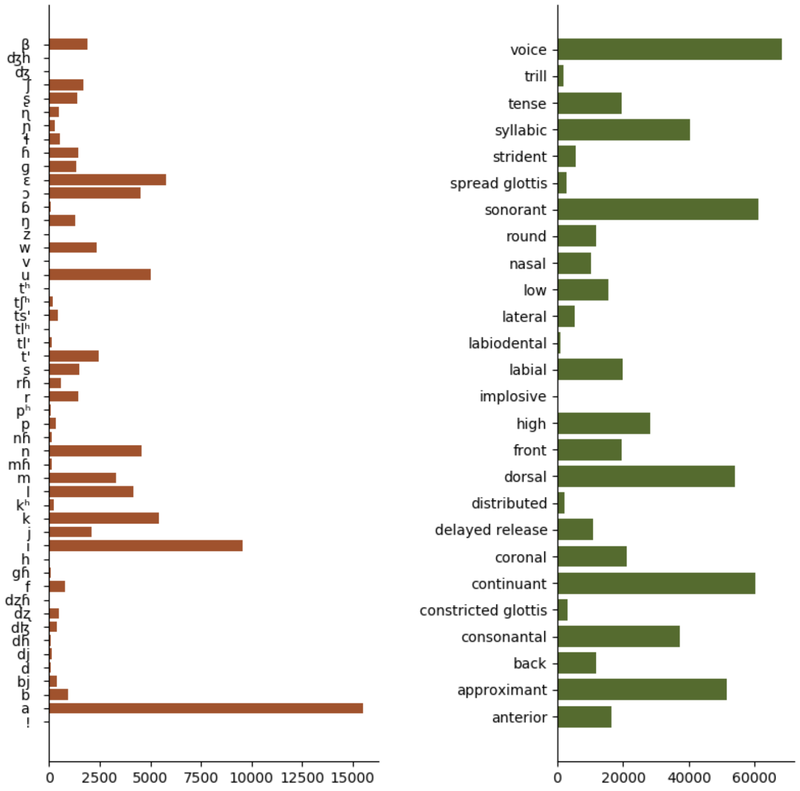
Fang Zheng, Guoliang Zhang, and Zhanjiang Song. 2001. Comparison of different implementations of MFCC. *Journal of Computer science and Technology*, 16(6):582–589.

Eberhard Zwicker. 1961. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248.

A Phonological feature definitions

We adopt the phonological feature definitions presented in Hayes (2011). For full exposition of the features and their motivations, we refer readers to the source. However, for convenience, we provide the following brief (and in some cases oversimplified) definitions based on Hayes (2011):

- **syllabic:** Vowels are [+syllabic], others are [-syllabic]
- **consonantal:** Vowels and glides are [-consonantal], others are [+consonantal]
- **approximant:** Vowels, liquids, and glides are [+approximant], others are [-approximant]

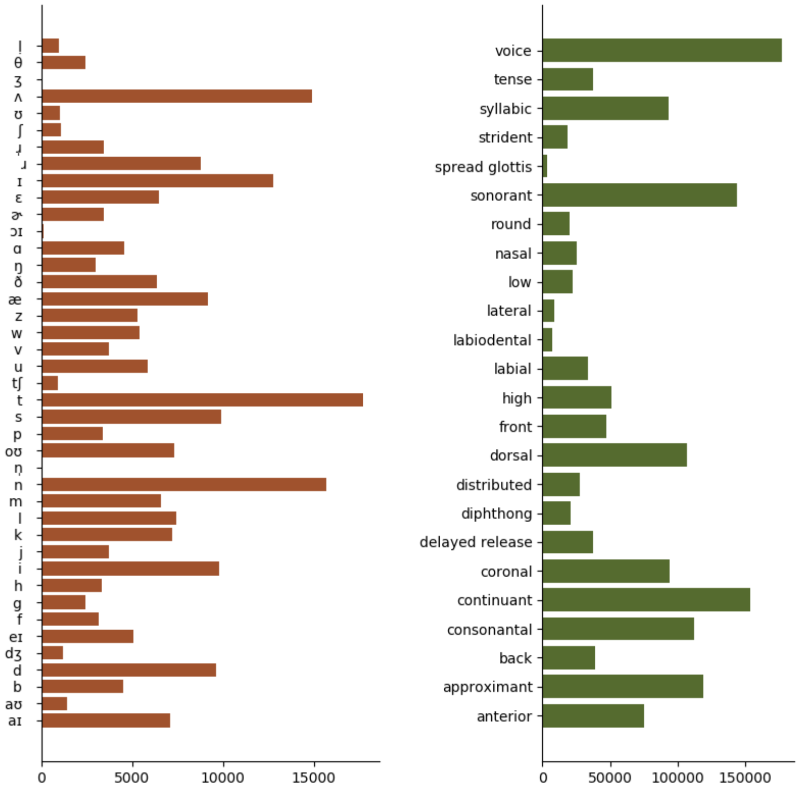


(a) Phonemes

(b) Features

Figure 3: Xitsonga phoneme and feature distributions.

- **sonorant:** Vowels, liquids, glides, and nasals are [+sonorant], others are [-sonorant]
- **continuant:** Stops and affricates are [-continuant], others are [+continuant]
- **delayed release:** Affricates and fricatives are [+delayed release], others are [-delayed release]
- **trill:** Trills are [+trill], others are [-trill]
- **front:** Front vowels and fronted velars are [+front], others are [-front]
- **back:** Back vowels and back velars are [+back], others are [-back]
- **high:** High vowels and velars are [+high], others are [-high]
- **low:** Low vowels and pharyngeals are [+low], others are [-low]
- **tense:** Tense vowels are [+tense], others are [-tense]
- **round:** Rounded vowels and rounded labial consonants are [+round], others are [-round]
- **nasal:** Nasal consonants and (contrastively) nasalized vowels are [+nasal], others are [-nasal]
- **labial:** Sounds articulated with the lips are [+labial], others are [-labial]
- **coronal:** Sounds articulated with the tongue blade/tip are [+coronal], others are [-coronal]
- **dorsal:** Sounds articulated with the tongue body are [+dorsal], others are [-dorsal]
- **anterior:** Coronals articulated at the alveolar ridge or forward are [+anterior], others are [-anterior]
- **distributed:** Coronals articulated with the tongue blade are [+distributed], others are [-distributed]
- **strident:** Sibilants (i.e. coronal fricatives and affricates) are [+strident], others are [-strident]
- **lateral:** Sounds with lateral oral closure (open at edges, like [l]) are [+lateral], others are [-lateral]
- **labiodental:** Sounds that are articulated by touching the lower lip to the upper teeth are [+labiodental], others are [-labiodental]



(a) Phonemes

(b) Features

Figure 4: English phoneme and feature distributions.

- **voice:** Voiced sounds are [+voice], others are [-voice]
- **spread glottis:** [h], [ɦ], and (contrastively) aspirated consonants are [+spread glottis], others are [-spread glottis]
- **constricted glottis:** Ejectives and glottal stops are [+constricted glottis], others are [-constricted glottis]
- **implosive:** Implosives are [+implosive], others are [-implosive]

B Xitsonga Phoneme Featurization

To the best of our knowledge, the gold Xitsonga phone transcriptions provided by the Zerospeech 2015 dataset use a non-standard pronunciation alphabet that is undocumented but isomorphic to the NCHLT transcription convention. In order to extract distinctive features for the Xitsonga phone labels, we hand-mapped the Zerospeech labels onto NCHLT labels by cross-referencing the Zerospeech phone sequences, the Zerospeech orthographic word sequences, and the NCHLT pronunciation dictionary, searching for systematic correspondences between Zerospeech and NCHLT transcription practices. Once the Zerospeech-to-NCHLT mapping was obtained, we used the International Phonetic Alphabet (IPA) phone labels provided by NCHLT to look up distinctive features in the Phonological CorpusTools (PCT) feature maps (Hall et al., 2016). Some IPA labels from NCHLT were not found in the PCT database, and for those we used the following featurization rules:

- **Consonants with palatal offglides:** We used the features associated with the non-offglide consonant and switched on the *approximant*, *dorsal*, *high*, *front*, and *tense* features.
- **Aspirated consonants:** We used the features associated with the non-aspirated consonant and switched on the *spread glottis* feature.
- **Ejective consonants:** We used the features associated with the non-ejective consonant and switched on the *constricted glottis* feature.
- **Voiceless alveolar lateral stops:** We used the features associated with voiceless alveolar lateral affricates and switched off the *de-*

layed release feature.

Our hand-made symbol correspondences and featurizations are distributed with this project's code repository.

C Phoneme and feature distributions

For reference, counts of phonemes and features by corpus are plotted in Figures 3 and 4. Note that the feature counts are generally larger because multiple features can be true of any one segment.