Acquiring language from speech by learning to remember and predict

Cory Shain
Ohio State University
shain.3@osu.edu

Micha Elsner
Ohio State University
elsner.14@osu.edu

Abstract

Classical accounts of child language learning invoke *memory* limits as a pressure to discover sparse, language-like representations of speech, while more recent proposals stress the importance of *prediction* for language learning. In this study, we propose a broad-coverage unsupervised neural network model to test memory and prediction as sources of signal by which children might acquire language directly from the perceptual stream. Our model embodies several likely properties of real-time human cognition: it is strictly incremental, it encodes speech into hierarchically organized labeled segments, it allows interactive top-down and bottom-up information flow, it attempts to model its own sequence of latent representations, and its objective function only recruits local signals that are plausibly supported by human working memory capacity. We show that much phonemic structure is learnable from unlabeled speech on the basis of these local signals. We further show that remembering the past and predicting the future both contribute to the linguistic content of acquired representations, and that these contributions are at least partially complementary.

1 Introduction

How children acquire language from the environment is one of the fundamental mysteries of cognitive science. Much theoretical, experimental, and computational research into this question has focused on acquiring abstractions over lower-order symbols, such acquiring morphemes from phoneme sequences or syntactic structures from word sequences (Chomsky, 1965; Gold, 1967; Elman, 1991; Saffran et al., 1996; Albright, 2002; Klein and Manning, 2004; Goldwater et al., 2009; Christodoulopoulos et al., 2012, *inter alia*). Children, however, do not get symbolic input; symbolic representations at any level of granularity constitute abstractions inferred from highly variable, noisy, and information-rich perceptual signals like audition and vision. This work joins a growing computational literature exploring the kinds of architectures and learning objectives that best support acquisition of linguistic representations directly from the speech signal without supervision (Versteegh et al., 2015; Dunbar et al., 2017). Such models can be used to test questions about language acquisition under more realistic assumptions about the input signal, especially to the extent that they reflect known constraints on human cognition (Shain and Elsner, 2019; Beguš, 2020).

This study uses computational modeling to examine two influential and possibly complementary ideas about how people learn abstract representations, including language, from data: learning to remember, and learning to predict. Both hypotheses have been advocated by prior work in language acquisition, cognitive neuroscience, and computational modeling, yet their relative contributions to language learning are not yet clear. Our model permits precise manipulation of memory and prediction pressures during acquisition, allowing direct comparison of these hypotheses.

In so doing, we implement several constraints on real-time language processing that have not been simultaneously present in prior modeling of this domain: (1) we jointly segment and label the speech signal without supervision; (2) the learning objective is applied incrementally during real-time processing using only locally available feedback; (3) the encoded signal is segmental, sparse, and hierarchically organized; (4) segments are represented featurally as patterns of activation, rather than discrete category symbols; and (5) the system is optimized by modeling its own state at multiple timescales, rather than by modeling the data alone.

Results show a systematic improvement along multiple measures of phoneme induction quality

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 195–214

https://doi.org/10.18653/v1/P17

from both learning to remember and learning to predict, suggesting that these two kinds of signals may play complementary roles during child language acquisition. The contributions of this work are as follows:

- We propose a novel deep neural encoder-decoder for unsupervised speech processing that is incremental, segmental, and useful for testing hypothesized cognitive constraints.

- We show empirically that memory-based and prediction-based signals contribute separately to the acquisition of linguistic regularities, simultaneously supporting two existing classes of theories about the learning pressures that underlie human language acquisition.

2 Background

2.1 Memory, Prediction, and Learning

Many proposals from the language acquisition literature appeal to memory pressures as a learning signal (Newport, 1990; Pinker, 1991; Carstairs-McCarthy, 1994; Rissanen and Ristad, 1994; Baddeley et al., 1998; Goldsmith, 2003; Yang, 2005, *inter alia*). For example, Baddeley et al. (1998) invoke constraints on working memory, arguing that because the speech signal is too rich to support full retention during real-time language processing (Baddeley and Hitch, 1974), infants are guided toward phonemic representations, which constitute an efficient encoding of that signal. Meanwhile, classical theories of language acquisition such as Newport (1990) and Pinker (1991) invoke constraints on long-term memory, arguing that linguistic regularities constitute compressed descriptions of the learner's input and that their discovery reduces the amount of information that must by idiosyncratically stored. Artificial language learning patterns in humans (Kersten and Earles, 2001) and recent computational modeling of the speech domain (e.g. Lee and Glass, 2012; Lee et al., 2015; Kamper et al., 2015; Elsner and Shain, 2017; Kamper et al., 2017a; Shain and Elsner, 2019) have supported a contribution from memory constraints to language learning. This position also aligns with an extensive computational neuroscience literature on *sparse coding*, which holds that biological neurons are tuned for memory-efficient representations of recent stimuli (Attneave, 1954; Olshausen and Field, 1996, 2004; Sheridan et al., 2017).

Nonetheless, debate exists about the role of memory in language learning. For example, Rohde and Plaut (1999) fail to replicate findings from Elman (1993) in favor of Newport (1990). In addition, Perfors (2012) fails to find evidence that memory bottlenecks encourage discovery of underlying linguistic regularities in adults and argues that such limitations only support language learning in concert with strong inductive priors. Furthermore, evidence suggests that mental representations during language processing preserve acoustic details over and above symbolic codes (Andruski et al., 1994; McMurray et al., 2002). Related work has called into question both the memory efficiency of human mental representations and the severity of long-term memory limits. For example, experimental evidence indicates that human mental representations contain redundant information, both of language (Baayen et al., 1997) and of other constructs such as logical relations (Piantadosi et al., 2016). In addition, recent estimates of mental storage requirements indicate that lexical information, especially semantics, already requires vastly more storage than e.g. phonemes and syntax, suggesting little added memory benefit from optimizing the efficiency with which regularities are stored (Mollica and Piantadosi, 2019). Finally, recent computational evidence linking memory bottlenecks to success in unsupervised speech processing has relied on storage of arbitrarily long acoustic sequences in their full detail in order to compute reconstruction losses (Kamper et al., 2015; Elsner and Shain, 2017). This design is inconsistent with known constraints on the storage duration ($<$ 1s) of unanalyzed acoustic traces in human working memory (Baddeley and Hitch, 1974; Cowan, 1984). It is thus not yet clear (1) how strongly memory pressures constrain mental representations of speech or (2) how much they encourage language learning.

Memory efficiency is not the only objective that can be constructed to learn abstractions over data without supervision. It has also been proposed that language learning may be driven by optimizing prediction of future input (Rohde and Plaut, 1999; Johnson et al., 2013; Phillips and Ehrenhofer, 2015; Apfelbaum and McMurray, 2017). This proposal aligns with an extensive neuroscience literature arguing that *predictive coding* for future inputs is a "canonical computation" of the human brain (Keller and Mrsic-Flogel, 2018) and may better characterize the tuning of biological neurons than

sparse coding (Singer et al., 2018), possibly because prediction affords advantages in critical tasks (Nijhawan, 1994) and may help organisms filter noise from the perceptual signal by focusing attention on features relevant to prediction (Bialek et al., 2001). Additional support for a role of prediction in language learning comes from the success of incremental *language models* in natural language processing, which optimize prediction of future words (Ney et al., 1994; Heafield et al., 2013; Jozefowicz et al., 2016; Radford et al., 2019). Language models support dramatic performance improvements in language processing tasks (Radford et al., 2019) and have been shown to both (1) acquire linguistic abstractions without direct supervision (Linzen et al., 2016) and (2) covary with human language comprehension measures (Frank and Bod, 2011; Goodkind and Bicknell, 2018; van Schijndel and Linzen, 2018). Finally, experimental evidence indicates that infants chunk the speech stream at points of low transition probability, suggesting that predictive signals are exploited to learn word-like units (Saffran et al., 1996).

We address these questions computationally by manipulating the presence or absence of memory and prediction pressures in the joint objective of an unsupervised incremental speech processing model, allowing us to quantify the contributions of these two hypothesized learning signals under realistic constraints on real-time processing.

2.2 Recurrent, Hierarchical, and Segmental Speech Processing in Humans

Artificial recurrent neural networks such as those employed here were initially proposed as algorithmic-level (Marr, 1982) models of activity in biological neural networks (Little, 1974; Hopfield, 1982), and subsequent studies support ubiquitous recurrence in the cortex (Harris and Mrsic-Flogel, 2013). In addition, influential theories of biological neural information processing argue that biological neural circuits integrate information at multiple hierarchically-organized timescales (Kiebel et al., 2008; Hasson et al., 2015; Norman-Haignere et al., 2020). Further neuroscientific evidence indicates that segmentation of the time dimension plays a critical role in human cognition, both in domain-general event processing (Zacks et al., 2001; Jensen, 2006, *inter alia*) and in speech processing specifically (Sanders and Neville, 2003; Cunillera et al., 2006, 2009; Kooijman et al., 2013; Lee and Cho, 2016, *inter alia*). Segmentation or "chunking" also plays a central role in several theories of language comprehension (Sanford and Sturt, 2002; Hale, 2006; Frank and Christiansen, 2018) and learning (Monaghan and Christiansen, 2010; McCauley and Christiansen, 2019). Our model incorporates these notions architecturally, with segment boundaries implemented by "detector neurons" that govern information flow between neural populations at larger and smaller timescales (Masquelier, 2018).

2.3 Modeling the Mental State

Many theories of linguistic structure posit multiple, hierarchically organized levels of representation (Chomsky, 1957; Goldsmith, 1976). Such theories predict the existence of abstractions over abstractions, latent structures that describe the distribution of other latent structures. This idea accords with recent theories of generalized Bayesian learning in biological agents, in which neural populations are thought to model the activity of other neural populations within their Markov blanket (Friston, 2010). The notion of learning through modeling other elements of the agent's own mental state has been exploited in symbolic computational models of language acquisition (Lee and Glass, 2012; Lee et al., 2015), but not in the context of artificial neural zero-resource speech models, which have so far derived their objective exclusively from the data (Kamper et al., 2017a; Elsner and Shain, 2017). Our approach incorporates this idea by optimizing higher layers to predict the sequence of activations at lower layers.

2.4 Related Computational Approaches

This work is part of a growing interest in unsupervised representation learning from raw speech, especially the Zerospeech 2015 (Versteegh et al., 2015) and 2017 (Dunbar et al., 2017) shared tasks and 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), as well as subsequent deep neural autoencoders (Van Den Oord et al., 2017; Chorowski et al., 2019) inspired by the WaveNet architecture (van den Oord et al., 2016). Much of this work concerns the discovery of word-like

units, while our analyses focus on learning at the phoneme level (see section 4.3).

A symbolic Bayesian framework for joint unsupervised phoneme segmentation and clustering is proposed by Lee and Glass (2012) and extended by Lee et al. (2015). Their system infers a Dirichlet process hidden Markov model to learn a symbolic sequential encoding of the speech stream. A disadvantage of this approach for the present research question is that the categorically distributed phone labels lack any notion of featural relatedness, contrary to widely held assumptions about natural language phonology (Clements, 1985). In addition, the learning signal derives from a next-frame prediction objective, making it difficult to use the model to factorially manipulate memory and prediction pressures. Another recent framework for unsupervised phone segmentation identifies boundaries at points of high surprisal in a frame-level language model (Michel et al., 2017). This approach does not generate segment encodings and cannot straightforwardly be used to test claims about the role of memory in language learning.

3 Model

Like many prior ANN zero-resource speech processing models (e.g. Kamper et al., 2015, 2017a; Elsner and Shain, 2017; Shain and Elsner, 2019), we employ an encoder-decoder framework. However, unlike previous approaches, our model decodes incrementally and hierarchically, with each layer decoding its inputs at their own timescale over a short window backward into the past and/or forward into the future. The model is thus required not only to describe the input signal (speech), but also its own sequence of latent representations (e.g. phones, words, etc.), much as people are implicitly thought to do in prior symbolic work on unsupervised language learning (Goldwater et al., 2009; Lee et al., 2015). Our encoder model closely follows Chung et al. (2017), and thus the primary technical contribution of this work lies in the cascaded incremental decoder and the layerwise incremental objective described below, both of which are designed to encourage repurposable segment representations based on locally available information. Although encodings are ultimately the quantity of interest in unsupervised encoder-decoder models, prior work has shown that decoder design can be a major determinant of acquired representations (McCoy et al., 2018, 2020). The overall design is schematized in Figure 1. Code is available at `https://github.com/coryshain/dnnseg`.

3.1 Encoder

Our encoder closely follows a hierarchical multiscale extension (HM-LSTM, Chung et al., 2017) of long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). The encoder consists of multiple LSTM layers linked by discrete boundary neurons that govern memory retention and information flow between layers. When a boundary neuron fires in layer l, it terminates a *segment*. Layer l then ejects its hidden state representation to layer $l+1$, receives top-down input from the hidden state of layer $l+1$, and resets its cell state (incremental memory) in order to process the next segment. The hidden state at a boundary thus constitutes a *label* for the segment terminating at that boundary, which is used to summarize the content of the segment when communicating with other layers. When the boundary neuron at layer l does not fire, layer $l+1$ is inert and simply copies its representation forward. As a result, higher layers track information at longer timescales than lower layers, and the segmentation behavior at l determines the input timescale at $l+1$. Each layer proceeds by segmenting and labeling its input signal at a timescale learned from data, resulting in a hierarchical sequence of labeled segments. As argued in Chung et al. (2017), this design enforces a trade-off between recurrent information (which is erased by segmentation) and top-down information (which is made available by segmentation).[1] Although the linguistic quality of discovered HM-LSTM segments is not systematically examined in the original proposal (Chung et al., 2017) and recent analysis has called it into question (Kádár et al., 2018), our results indicate that HM-LSTMs can discover segmental structure from speech, at least at the phonemic level.

3.2 Decoder

The decoder consists of two multi-layer attentional sequence-to-sequence (seq2seq) LSTMs with L layers each, one backward-directional (memory) and one forward directional (prediction). The LSTMs respectively decode the B previous input segment labels and the F following input segment labels given an encoder representation at layer l and

[1] See Appendix A for definition of the encoder and B for comparison to Chung et al. (2017).

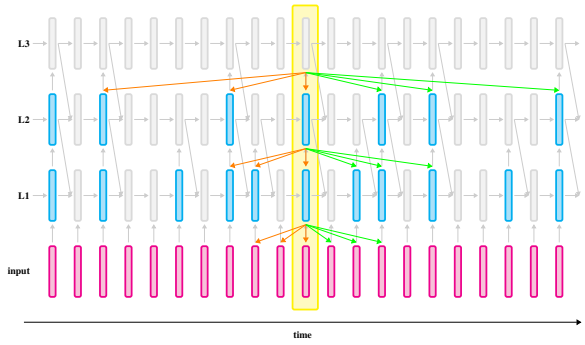

Figure 1: *Incremental layerwise encoder-decoder framework.* Shown here with 3 layers and a forward/backward window size of 3. Segment boundaries are shown in cyan. Gray arrows indicate information flow through the encoder, as governed by the boundary decisions. Colored arrows indicate information flow from encodings to decoder targets in the backward (orange) and forward (green) directions, starting from the encoded timestep at the center of the figure.

time t. In addition, the predicted sequence from the layer above serves as attention values to inform decoding at the current layer, at a timescale determined by the segmentation patterns of the current layer and the layer below. The internal behavior of the decoder is thus tightly coupled with the segmental behavior of the encoder, providing direct feedback into the encoder decisions. In addition, the label sequence of the decoder at all layers must support both (1) decoding of the perceptual signal (the data), since top-down connections allow higher-level representations to inform lower-level ones, and (2) decoding of lower-level state sequences.[2]

3.3 Objective

We employ an incremental layerwise objective that both *reconstructs* backward and *predicts* forward from time t at layer l over the segment labels from layer $l-1$ at a timescale defined by the segmentation behavior of layer $l-1$. Thus only the first layer decodes at the timescale of the data; higher layers l decode at the timescale of $l-1$, and representations associated with non-boundaries in $l-1$ are ignored by the objective. The objective scans incrementally over the time dimension and imposes a forward and backward cost at every segment boundary identified by layer $l-1$. As a result, the first layer ("phonemes") is responsible for incrementally decoding the local past and future realization of the acoustic stream, the second layer is responsible for decoding the local past and future realization of the "phoneme" sequence, etc.[3]

Although it is possible to backpropagate into the decoding targets (i.e. encoder representations) at higher layers, thereby encouraging the encoder to discover more predictable segment sequences, we found in practice that doing so resulted in a form of mode collapse where labels became insensitive to the data and converged to a single value for all timesteps. For this reason, we stop the gradients into decoding targets and backpropagate only into the decoder predictions. Thus, the objective encourages encodings at higher layers to change to better predict structures at lower layers— but does not alter the representations at those lower layers to make them more uniform and therefore easier to predict.

4 Experimental Design

We assess the contribution of memory and prediction pressures to phoneme acquisition by (1) manipulating these pressures on models exposed to speech data from two unrelated languages (Xitsonga and English) and (2) evaluating the effect of these manipulations on multiple measures of phoneme induction quality.

4.1 Data

We use the Zerospeech 2015 (Versteegh et al., 2015) challenge data in English and Xitsonga, a Bantu language spoken in South Africa. The Xitsonga data come from the NCHLT corpus (De Vries et al., 2014) and contain 2h29m07s of read speech from 24 speakers. The English data come from the Buckeye Corpus (Pitt et al., 2005) and contain 4h59m05s of spontaneous speech from 12 speakers. For English, we additionally include the official development set in training, which contains 1h39m45s of spontaneous speech from 2 speakers, also from the Buckeye Corpus. English development set performance was used for model development and tuning, but the development set is not included in the evaluations presented here. Xitsonga lacks a development set, so designs selected on the English development set are applied directly to Xitsonga for evaluation. Before fitting, we convert the source audio files into a cochleagram-based spectral representation that approximates the signal generated by the human auditory system (McDermott and Simoncelli, 2011).[4]

[2] See Appendix C for definition of the decoder.

[3] See Appendix D for definition of the objective.

[4] See Appendix F for full preprocessing details.

4.2 Experimental Manipulations

We seek to assess the contribution of both memory and prediction pressures to the content of model representations. To this end, our principal manipulations are the backward ($B \in \{0, 5, 25, 50\}$) and forward ($F \in \{0, 1, 5, 10\}$) window lengths used by the decoder, which respectively impose a pressure to efficiently remember and accurately predict. Note that the condition $B = 0$, $F = 0$ (no reconstruction or prediction) is not well defined because the objective is 0 at any parameterization and thus has no gradient, and we therefore exclude it from consideration in these results. In addition, we manipulate the number of encoder layers ($L \in \{2, 3, 4\}$). This is because it is unclear *a priori* which layers of the encoder are expected to correspond to quantities of interest like phonemes or words, since the representations are unsupervised and the model could additionally or instead discover e.g. subphonemic, morphemic, phrasal, intonational, and other kinds of structures. Although detection of these and other levels of linguistic representation is of interest and is the target of future work, the annotations provided by the Zerospeech 2015 data support phoneme-level and word-level analyses only, and we concentrate our evaluation there. Varying the number of layers allows us to investigate which layers emergently discover more phoneme-like units, and under what conditions.

4.3 Evaluation

Because our model generates a segmental encoding of the speech signal, we apply two classes of evaluation in this study: phoneme segmentation and phoneme-level probing classification. The segmentation evaluation measures the degree of correspondence between the model-generated segment boundaries and expert-annotated phoneme boundaries, using a boundary F-measure which assigns a true positive for up to one predicted boundary that falls within some tolerance of each gold boundary, false positives for all other predicted boundaries, and false negatives for all gold boundaries that lack a predicted boundary within the tolerance. Following Lee and Glass (2012), we use a tolerance of 20ms. The classification evaluation measures the amount of signal in model-generated encodings as to (1) the true identity of the phoneme being encoded and (2) the cluster of phonological features associated with that phoneme (Hayes, 2011). Following e.g. Shain and Elsner (2019) and Chrupała et al. (2020), we do so using probing classifiers. In particular, for each layer of each model's encoder, we fit linear classifiers to (1) the phoneme labels and (2) the phonological feature labels associated with the gold phoneme segment corresponding to each phone boundary. We extract the gold and predicted phoneme encodings at the human-annotated phoneme boundaries, regardless of whether the model segmented at that location. This supports direct comparison of metrics across models, since the set of evaluated segments is held constant. Phonological features are extracted at the same timepoints, following the procedure described in Shain and Elsner (2019).[5]

Although our model is designed to support joint discovery of multiple layers of representation, we find empirically that no model appreciably improves at any layer in word boundary F-score over a baseline that segments only at the ends of voice activity regions, and qualitative inspection does not indicate systematic correspondence to an unannotated level of representation such as syllables, morphemes, or intonational units. Despite differences in segmentation rate, and thereby in word boundary precision-recall trade-off, models generally converge on similar (low) word boundary F-scores, and thus our manipulations are not informative about word learning. Probe-based classification metrics are not well suited to word-level evaluation due to the size of the vocabulary. Though human speech processing involves units between the phoneme and word level, detailed analysis of such units is difficult due to the lack of annotation in the corpus. We believe poor word discovery at higher layers may be due in part to the fact that non-initial layers have both a non-stationary objective (the evolving representations of the layer below) and slower learning dynamics, perhaps making it difficult for these layers to "catch up" with moving targets (Ioffe and Szegedy, 2015). We leave exploration of possible remedies to future research and focus here only on the phoneme level.

While it is *a priori* unclear which layer of the encoder is expected to encode phonemes (for example, the initial layers may encode sub-phonemic units), we find systematically better phoneme segmentation and classification performance in the first layer of the network. For simplicity, we therefore only present metrics from this layer.

In addition to reporting raw model performance,

[5] See Appendix E for probe implementation details.

Model	English			Xitsonga		
	Bd	Pc	Fc	Bd	Pc	Fc
Full	**65.3**	**22.9**	**49.3**	39.3	**28.6**	**53.8**
Baseline U	30.4	12.3	42.2	22.1	15.4	46.2
Baseline X	52.4	20.5	47.1	**44.8**	27.8	53.2

Table 1: F-measures for boundary discovery (Bd), phoneme classification (Pc), and phonological feature classification (Fc), using $B = 25$, $F = 1$, $L = 3$.

we report performance improvements from each model relative to (1) baseline U (*untrained*), an architecturally matched model left at random initialization (Chrupała et al., 2020), and (2) baseline X (*cross-language*), the architecturally matched model trained on the opposite language.[6] These two baselines quantify different contributions of the acquisition process. Baseline U quantifies architectural inductive bias: how well does the architecture alone guide linguistic representations, without learning? Baseline X quantifies modality inductive bias: how well does general knowledge of human speech guide linguistic representations, without exposure to the target language? Improvement against either of these baselines supports language learning from experience, over and above any prior knowledge that might more efficiently be innately encoded.[7]

5 Results and Discussion

Boundary and macro-averaged phoneme and feature classification F-measures from the best-performing configuration on the English development set ($B = 25$, $F = 1$, $L = 3$) are given in Table 1. English boundary performance ($F = 65.3$) approaches previously reported unsupervised phoneme segmentation scores on different and therefore not directly comparable datasets (Lee and Glass, 2012; Michel et al., 2017, both around

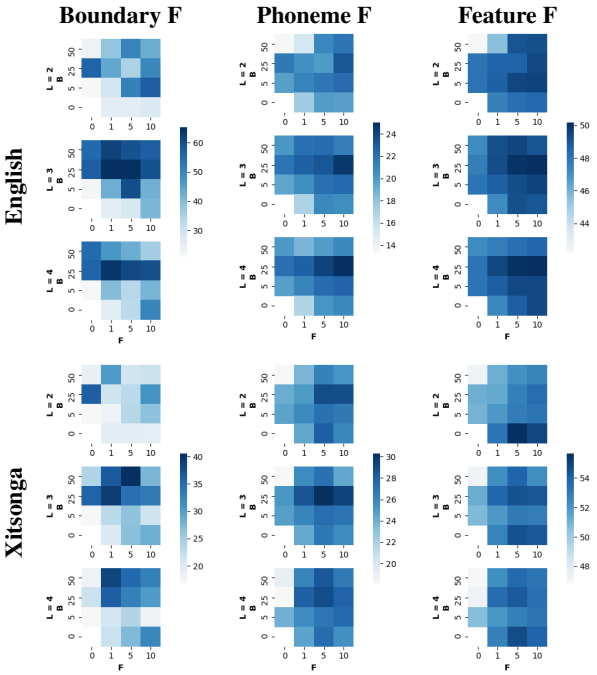

Figure 2: **Phoneme acquisition scores.** F-measures for boundary detection (left), phoneme classification (center), and phonological feature classification (right).

$F = 75$). The overall segmentation performance in Xitsonga is considerably worse than that of English, consistent with prior evidence that word segmentation in the Zerospeech 2015 Xitsonga partition is harder than English (e.g. Kamper et al., 2017a). By contrast, classification metrics in Xitsonga are better than in English, which is again consistent with prior findings of stronger unsupervised classification performance in Xitsonga (Shain and Elsner, 2019).

The difference in relative performance between segmentation and classification in the two languages could be due in part to differences in register: the English data is spontaneously produced, while the Xitsonga data is read speech. Longer average phoneme duration (100ms vs 70ms) and cleaner articulations in Xitsonga could plausibly give rise to this asymmetry, and further investigation is left to future work. The model substantially outperforms the untrained baseline (U) on all metrics and outperforms the cross-language baseline (X) on all metrics but boundary F in Xitsonga, which could be due in part to the larger size of the English-language training set. Results therefore indicate that the reconstruction and prediction objectives have contributed to unsupervised discovery of phonemic patterns in both languages.

Segmentation and macro-averaged classification

[6]For Xitsonga, baseline X is the architecturally matched English-trained model. For English, baseline X is the architecturally matched Xitsonga-trained model.

[7]We do not evaluate directly against a previous state of the art because no state of the art exists for unsupervised phoneme segmentation and classification in the Zerospeech 2015 data. A previous model that performed the same task (Lee and Glass, 2012) achieved an average boundary F-score of 76.1 on a different dataset that used a different boundary annotation standard (automatic forced alignment instead of human annotation). To our knowledge, the dataset is no longer publicly available. A recent segmentation-only model (Michel et al., 2017) achieved a boundary F of 75 on the TIMIT dataset (Fisher et al., 1986). However, because TIMIT is restricted to 10 unique utterances of English, we believe Zerospeech 2015, which contains more linguistically diverse speech from two unrelated languages, is a better dataset for investigating language acquisition patterns.

F-measures by language and experimental condition are given in Figure 2. Results show a contribution of both memory ($B > 0$) and prediction ($F > 0$), with a similar distribution of relative performance between the two languages, supporting the existence of language-general influences of prediction and memory on phoneme learning.

As shown in Figure 2, models without memory pressures ($B = 0$) find substantially worse boundaries than models with memory pressures. There also appears to be a ceiling effect of backward reconstruction size, with a jump in performance at $B = 25$ but no systematic improvement at $B = 50$. Importantly, at layer 1, $B = 25$ covers a 250ms interval, which falls within even conservative estimates of the storage duration of unanalyzed auditory traces in humans (Cowan, 1984). The $B = 25$ objective could therefore plausibly be used during online speech processing. Prediction pressures also support discovery of phoneme boundaries, as shown by the generally worse boundary performance of $F = 0$ vs. $F > 0$ in both languages.

In addition, Figure 2 shows that memory and prediction both modulate phoneme classification performance, with a roughly convex performance surface around a peak at $B = 25$, $F = 10$ for English and $B = 25$, $F = 5$ for Xitsonga. A similar peak emerges in the feature classification results for English, along with a local feature classification peak in Xitsonga for $L > 2$. A 250ms auditory memory window thus supports both phoneme segmentation and classification in our models, with additional benefits from predicting over short intervals (Singer et al., 2018). For feature classification, the primary determinant of performance across languages is the prediction objective, with performance generally increasing up to $F = 5$. There is also an effect of encoder depth in these results, such that encoders with more layers ($L > 2$) tend to perform better across metrics, despite the fact that all metrics reflect performance at the first layer. This result supports a contribution of multiscale modeling, even if the segmentation behavior at higher layers does not clearly correspond to a theory-driven level of representation (see section 4.3).

Figure 3 reports performance differences by metric against baseline U (untrained). Training yields consistent and often substantial improvements across metrics in multiscale ($L > 0$) encoders with both memory ($B > 0$) and prediction ($F > 0$) pressures, but can fail to improve in the

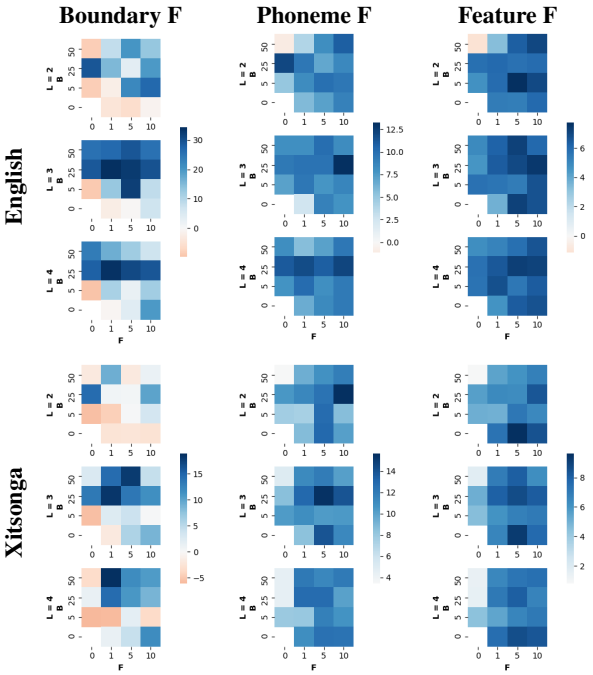

Figure 3: **Effect of learning.** Change in F-measure by metric over baseline U.

absence of these characteristics. Memory and prediction therefore modulate not only absolute performance, but also the utility of language experience.

Figure 4 reports performance differences by metric against baseline X (cross-language). English segmentation is substantially helped by experience with (i.e. training on) English, especially under strong memory pressures. However, Xitsonga segmentation is generally worse for the Xitsonga-trained model than the English-trained one. This might be due to the fact that the English training set is larger, and/or to low overall levels of segmentation performance in Xitsonga. While we leave further investigation of this exception to future work, the classification metrics still show a clear benefit of in-domain training in both languages, but only in the presence of prediction pressures.

The baseline X results bear on the degree to which speech processing patterns can plausibly be innately specified. Although the set of phonological categories and features are classically regarded as universal (Chomsky and Halle, 1968; Clements, 1985), it is well known that the "same" phonological abstraction (e.g. voicing) can be phonetically cached out in different ways depending on the language (e.g. Gordon and Ladefoged, 2001; Gordon et al., 2002). Our results suggest that, at least between Xitsonga and English, this variation is both (1) constrained enough to permit recognition of

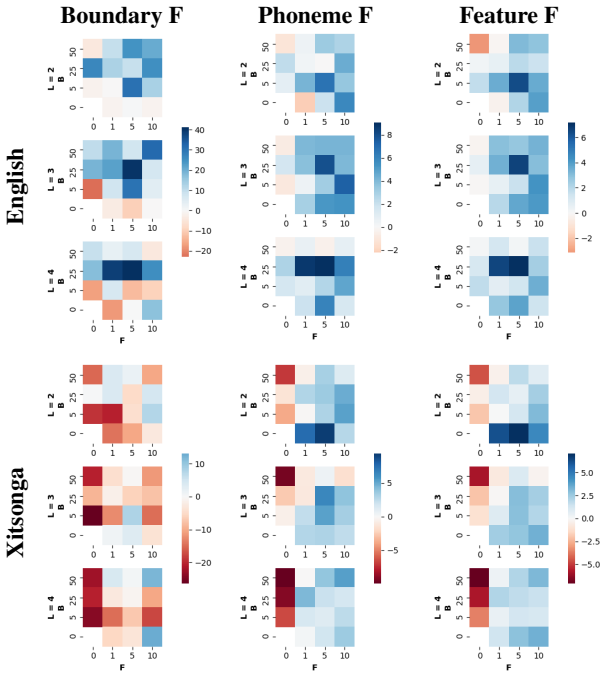

Figure 4: **Effect of language.** Change in F-measure by metric over baseline X.

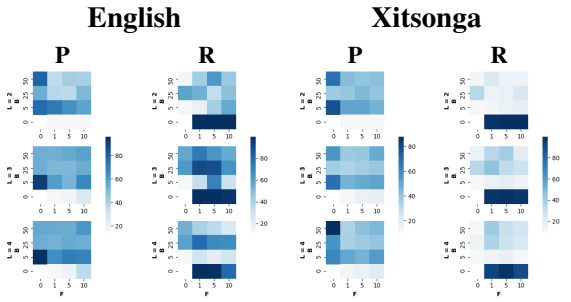

Figure 5: **Boundary P/R trade-off.** Boundary precision (left) and recall (right) by experimental configuration in Xitsonga and English.

non-trivial patterns from speech in other languages on the basis of general, possibly innate processing biases, and (2) substantial enough to give rise to a benefit of direct experience with the target language, even for language-general constructs like phoneme categories and phonological features.

We use linear regression on the combined metrics to quantitatively evaluate the contribution of both memory and prediction pressures to phoneme acquisition. Results show significant positive contributions to acquisition from memory pressures ($p = 0.006$), prediction pressures ($p < 0.001$), and multiscale encoding ($p < 0.001$).[8]

The boundary precision/recall trade-off illuminates the mechanisms by which memory and prediction pressures affect learning (Figure 5). Without memory pressures ($B = 0$), segmentation rates are high, resulting in high recall and low precision. Introducing memory pressures ($B > 0$) slows the segmentation rate, resulting in a more balanced P/R trade-off. Without prediction pressures ($F = 0$), segmentation rates are generally low, resulting in higher precision and lower recall. Introducing prediction pressures ($F > 0$) increases the segmentation rate, again resulting in a more balanced trade-off. To understand this pattern, recall that a boundary in our model represents both a cost (flushing the memory cell) and a benefit (injecting top-down feedback). The cost of forgetting is plausibly greater for reconstruction than prediction, since only the current layer has had direct access to the sequence of reconstruction targets. By contrast, the benefit of top-down feedback is plausibly greater for prediction than reconstruction, since the prediction can condition on contextual representations at multiple timescales. In our segmental model of speech processing, the objectives therefore induce countervailing biases that boost signal for phonological constructs, supporting their joint influence on phoneme acquisition from speech.

6 Conclusion

We proposed an unsupervised deep neural model of speech processing that is incremental, segmental, and optimized by local feedback. We manipulated the model's objective function in order to investigate prior hypotheses about the role in human language acquisition of memory constraints on the one hand and predictive processing on the other. Results support a role for both memory and prediction pressures for acquiring phonemes from speech. Both objectives inform the model's segmentation behavior and the content of its segment encodings. In addition, results suggest that these two mechanisms coordinate to support phoneme discovery by introducing countervailing pressures toward retention of previously encountered signals (memory) and consultation of top-down signals (prediction).

Acknowledgements

We thank Aren Jansen and the Clippers discussion group at Ohio State for providing valuable feedback on this research. This work was funded in part by a Google Faculty Research Award to Micha Elsner.

[8] See Appendix G for details.

References

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*.

Adam Albright. 2002. Islands of reliability for regular morphology: Evidence from Italian. *Language*, pages 684–709.

Jean E Andruski, Sheila E Blumstein, and Martha Burton. 1994. The effect of subphonetic differences on lexical access. *Cognition*, 52(3):163–187.

T K Ansari, Rajath Kumar, Sonali Singh, and Sriram Ganapathy. 2017a. Deep learning methods for unsupervised acoustic modeling—Leap 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.

Keith S Apfelbaum and Bob McMurray. 2017. Learning during processing: Word learning doesn't wait for word recognition to finish. *Cognitive science*, 41:706–747.

Fred Attneave. 1954. Some informational aspects of visual perception. *Psychological review*, 61(3):183.

R Harald Baayen, Ton Dijkstra, and Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37(1):94–117.

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*.

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*.

Gašper Beguš. 2020. Modeling unsupervised phonetic and phonological learning in Generative Adversarial Phonology. *Proceedings of the Society for Computation in Linguistics*, 3(1):138–148.

Pascal Belin, Robert J Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. 2000. Voice-selective areas in human auditory cortex. *Nature*, 403(6767):309–312.

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

William Bialek, Ilya Nemenman, and Naftali Tishby. 2001. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463.

Anna Blasi, Evelyne Mercure, Sarah Lloyd-Fox, Alex Thomson, Michael Brammer, Disa Sauter, Quinton Deeley, Gareth J Barker, Ville Renvall, Sean Deoni, and others. 2011. Early specialization for voice and emotion processing in the infant brain. *Current biology*, 21(14):1220–1224.

Guy J Brown and Martin Cooke. 1994. Computational auditory scene analysis. *Computer speech and language*, 8(4):297–336.

Andrew Carstairs-McCarthy. 1994. Inflection classes, gender, and the principle of contrast. *Language*, pages 737–788.

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.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

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

Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron van den Oord. 2019. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2012. Turning the pipeline into a loop: Iterated unsupervised dependency parsing and PoS induction. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 96–99.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. 2020. Analyzing analytical methods: The case of phonology in neural models of spoken language. *arXiv preprint arXiv:2004.07070*.

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

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

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

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*.

Nelson Cowan. 1984. On short and long auditory stores. *Psychological bulletin*, 96(2):341.

Toni Cunillera, Estela Càmara, Juan M Toro, Josep Marco-Pallares, Nuria Sebastián-Galles, Hector Ortiz, Jesús Pujol, and Antoni Rodr\'\iguez-Fornells. 2009. Time course and functional neuroanatomy of speech segmentation in adults. *Neuroimage*, 48(3):541–553.

Toni Cunillera, Juan M Toro, Nuria Sebastián-Gallés, and Antoni Rodr\'\iguez-Fornells. 2006. The effects of stress and statistical cues on continuous speech segmentation: an event-related brain potential study. *Brain Research*, 1123(1):168–178.

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.

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.

Jeffrey L Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

Ryan Eloff, André Nortje, Benjamin van Niekerk, Avashna Govender, Leanne Nortje, Arnu Pretorius, Elan van Biljon, Ewald van der Westhuizen, Lisa van Staden, and Herman Kamper. 2019. Unsupervised Acoustic Unit Discovery for Speech Synthesis Using Discrete Latent-Variable Neural Networks. *Proc. Interspeech 2019*, pages 1103–1107.

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.

Shirley Fecteau, Jorge L Armony, Yves Joanette, and Pascal Belin. 2005. Sensitivity to voice in human prefrontal cortex. *Journal of Neurophysiology*, 94(3):2251–2254.

William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall. 1986. The DARPA Speech Recognition Research Database: Specifications and Status. In *Proceedings of DARPA Workshop on Speech Recognition*, pages 93–99.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834.

Stefan L Frank and Morten H Christiansen. 2018. Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*, 33(9):1213–1218.

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Mark E. Gold. 1967. Language Identification in the Limit. *Information and Control*, (10):447–474.

John Goldsmith. 1976. *Autosegmental phonology*. Ph.D. thesis, MIT Press London.

John Goldsmith. 2003. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–189.

Sharon Goldwater, Thomas Griffiths, and Mark Johnson. 2009. A {Bayesian} framework for word segmentation: {Exploring} the effects of context. *Cognition*, 112:21–54.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18.

Matthew Gordon, Paul Barthmaier, and Kathy Sands. 2002. A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, pages 141–174.

Matthew Gordon and Peter Ladefoged. 2001. Phonation types: a cross-linguistic overview. *Journal of phonetics*, 29(4):383–406.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

Kenneth D Harris and Thomas D Mrsic-Flogel. 2013. Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58.

Uri Hasson, Janice Chen, and Christopher J Honey. 2015. Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, 19(6):304–313.

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

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

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.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

John J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

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.

Ole Jensen. 2006. Maintenance of multiple working memory items by temporal segmentation. *Neuroscience*, 139(1):237–249.

Matt A Johnson, Nicholas B Turk-Browne, and Adele E Goldberg. 2013. Prediction plays a key role in language development as well as processing. *Behavioral and Brain Sciences*, 36(4):360.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Ákos Kádár, Marc-Alexandre Côté, Grzegorz Chrupała, and Afra Alishahi. 2018. Revisiting the Hierarchical Multiscale LSTM. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3215–3227.

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.

Georg B Keller and Thomas D Mrsic-Flogel. 2018. Predictive Processing: A Canonical Cortical Computation. *Neuron*, 100(2):424–435.

Alan W Kersten and Julie L Earles. 2001. Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, 44(2):250–273.

Stefan J Kiebel, Jean Daunizeau, and Karl J Friston. 2008. A hierarchy of time-scales and the brain. *PLoS Comput Biol*, 4(11):e1000209.

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

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, volume 1, pages 478–485.

Valesca Kooijman, Caroline Junge, Elizabeth K Johnson, Peter Hagoort, and Anne Cutler. 2013. Predictive brain signals of linguistic development. *Frontiers in psychology*, 4:25.

Byeongwook Lee and Kwang-Hyun Cho. 2016. Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. *Scientific reports*, 6:37647.

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.

Chia-ying Lee, Timothy J O'Donnell, and James Glass. 2015. Unsupervised Lexicon Discovery from Acoustic Input. In *Transactions of the Association for Computational Linguistics*, volume 3, pages 389–403.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

William A Little. 1974. The existence of persistent states in the brain. *Mathematical biosciences*, 19(1-2):101–120.

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*.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

Timothée Masquelier. 2018. STDP allows close-to-optimal spatiotemporal spike pattern detection by single coincidence detector neurons. *Neuroscience*, 389:133–140.

Stewart M McCauley and Morten H Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological review*, 126(1):1.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2093–2098.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Josh H McDermott and Eero P Simoncelli. 2011. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940.

Bob McMurray, Michael K Tanenhaus, and Richard N Aslin. 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2):B33–B42.

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

Paul Michel, Okko Rasanen, Roland Thiollière, and Emmanuel Dupoux. 2017. Blind Phoneme Segmentation With Temporal Prediction Errors. In *Proceedings of ACL 2017, Student Research Workshop*, pages 62–68.

Francis Mollica and Steven T Piantadosi. 2019. Humans store about 1.5 megabytes of information during language acquisition. *Royal Society open science*, 6(3):181393.

Padraic Monaghan and Morten H Christiansen. 2010. Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of child language*, 37(3):545–564.

Elissa Newport. 1990. Maturational constraints on language learning. *Cognitive Science*, 14:11–28.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1):1–38.

R Nijhawan. 1994. Motion extrapolation in catching. *Nature*, 370(6487):256.

Sam V Norman-Haignere, Laura K Long, Orrin Devinsky, Werner Doyle, Ifeoma Irobunda, Edward Merricks, Neil A Feldstein, Guy M V McKhann, Catherine Schevon, Adeen Flinker, and Nima Mesgarani. 2020. Hierarchical integration across multiple timescales in human auditory cortex. *bioRxiv*.

Bruno A Olshausen and David J Field. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607.

Bruno A Olshausen and David J Field. 2004. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*, page 125.

Fabian Pedregosa, Gaël 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.

Amy Perfors. 2012. When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, 67(4):486–506.

Cyril R Pernet, Phil McAleer, Marianne Latinus, Krzysztof J Gorgolewski, Ian Charest, Patricia E G Bestelmeyer, Rebecca H Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, and others.

2015. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, 119:164–174.

Colin Phillips and Lara Ehrenhofer. 2015. The role of language processing in language acquisition. *Linguistic approaches to bilingualism*, 5(4):409–453.

Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. 2016. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392.

Steven Pinker. 1991. Rules of language. *Science*, 253(5019):530–535.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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*.

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*.

Jorma Rissanen and Eric Sven Ristad. 1994. Language acquisition in the MDL framework. *Language computations*, pages 149–166.

Douglas L T Rohde and David C Plaut. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Lisa D Sanders and Helen J Neville. 2003. An ERP study of continuous speech processing: I. Segmentation, semantics, and syntax in native speakers. *Cognitive Brain Research*, 15(3):228–240.

Anthony J Sanford and Patrick Sturt. 2002. Depth of processing in language comprehension: Not noticing the evidence. *Trends in cognitive sciences*, 6(9):382–386.

Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Marten van Schijndel and Tal Linzen. 2018. A Neural Model of Adaptation in Reading. In *EMNLP 2018*, pages 4704–4710.

Cory Shain and Micha Elsner. 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)*, pages 69–85.

Patrick M Sheridan, Fuxi Cai, Chao Du, Wen Ma, Zhengya Zhang, and Wei D Lu. 2017. Sparse coding with memristor networks. *Nature nanotechnology*, 12(8):784.

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.

Aaron Van Den Oord, Oriol Vinyals, and others. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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*.

Charles Yang. 2005. On productivity. *Linguistic variation yearbook*, 5(1):265–302.

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.

Jeffrey M Zacks, Todd S Braver, Margaret A Sheridan, David I Donaldson, Abraham Z Snyder, John M Ollinger, Randy L Buckner, and Marcus E Raichle.

2001. Human brain activity time-locked to perceptual event boundaries. *Nature neuroscience*, 4(6):651–655.

Pavel Zahorik and Frederic L Wightman. 2001. Loudness constancy with varying sound source distance. *Nature neuroscience*, 4(1):78–83.

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

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560.

A Encoder Definition

Given encoder inputs $\mathbf{x}_t^{\mathrm{e}} \in \mathbb{R}^{D_0}$ for $0 < t \leq T$, an encoder with L layers is a recurrent neural network that computes its D_l-dimensional hidden state $\mathbf{h}_t^{\mathrm{e}(l)} \in \mathbb{R}^{D_l}$ at timestep t and layer l from the hidden state below $\mathbf{h}_t^{\mathrm{e}(l-1)}$ (bottom-up connection), the previous hidden state $\mathbf{h}_{t-1}^{\mathrm{e}(l)}$ (recurrent connection), and the previous hidden state above $\mathbf{h}_{t-1}^{e(l+1)}$, where the layer zeroth state $\mathbf{h}_t^{\mathrm{e}(0)} \in \mathbb{R}^{D_0}$ is the data $\mathbf{x}_t^{\mathrm{e}}$. The hidden state $\mathbf{h}_t^{\mathrm{e}(l)}$ serves as a *label* at layer l and timestep t. Information flow between these layers is governed by a discrete boundary neuron $z_t^{(l)} \in \{0, 1\}$ at each layer. Let $t^{(l)\prime}$ be the location of the most recent segment boundary preceding time t in layer l:

$$t^{(l)\prime} \stackrel{\mathrm{def}}{=} \tau \mid \Big[(z_\tau^{(l)} = 1) \wedge \Big(z_{\tau'}^{(l)} = 0,\ \forall \tau' \in \{\tau+1, \ldots, t-1\}\Big)\Big] \quad \forall \tau \in \{1, \ldots, t-1\} \tag{A1}$$

Let $f^{\mathrm{f}(l)}(a, b)$ be a filter function dropping labels of l at non-boundaries between timepoints a and b:

$$f^{\mathrm{f}(l)}(a, b) \stackrel{\mathrm{def}}{=} \mathbf{h}_\tau^{\mathrm{e}(l)} \mid \Big(z_\tau^{(l)} = 1\Big),\ \forall \tau \in \{a, \ldots, b\} \tag{A2}$$

A *segment* $\mathbf{S}_t^{(l)}$ at layer l and timestep t is defined as:

$$\mathbf{S}_t^{(l)} \stackrel{\mathrm{def}}{=} f^{\mathrm{f}(l-1)}(t^{(l)\prime} + 1, t) \tag{A3}$$

In other words, the segment $\mathbf{S}_t^{(l)}$ consists of the sequence of segment labels from layer $l-1$ at boundaries from $l-1$ that intervene since the last segment boundary at l.

The bottom-up, recurrent, and top-down inputs are respectively linearly transformed into vectors in $\mathbf{s}^{\mathrm{b}(l)}, \mathbf{s}^{\mathrm{r}(l)}, \mathbf{s}^{\mathrm{t}(l)} \in \mathbb{R}^{4D_l+1}$ using weight matrices $\mathbf{W}_i^j$ mapping from layer i to layer j, and masked using the boundary decisions z:

$$\mathbf{s}_t^{\mathrm{b}(l)} \stackrel{\mathrm{def}}{=} z_t^{(l-1)} \mathbf{W}_{l-1}^l \mathbf{h}_t^{\mathrm{e}(l-1)} \tag{A4}$$

$$\mathbf{s}_t^{\mathrm{r}(l)} \stackrel{\mathrm{def}}{=} \mathbf{W}_l^l \begin{pmatrix} \mathbf{h}_t^{\mathrm{e}(l)} \\ \mathbf{h}_t^{\mathrm{e}(l)\prime} \\ n_t^{(l)} \end{pmatrix} \tag{A5}$$

$$\mathbf{s}_t^{\mathrm{t}(l)} \stackrel{\mathrm{def}}{=} z_{t-1}^{(l)} \mathbf{W}_{l+1}^l \mathbf{h}_t^{\mathrm{e}(l+1)} \tag{A6}$$

where $\mathbf{h}_t^{\mathrm{e}(l)\prime}$ records the label at the preceding segment boundary $\mathbf{h}_{t^{(l)\prime}}^{\mathrm{e}(l)}$ and $n_t^{(l)}$ is the number of timesteps since the preceding segment boundary at l. We pass this additional information into the recurrent connection to relieve pressure on the cell state to encode it. These vectors are summed together with a bias $\mathbf{b}^{(l)}$ to create a vector of preactivations $\mathbf{s}^{(l)}$, normalized by the boundary decisions so that the weights on active connections sum to 1:

$$\mathbf{s}_t^{(l)} \stackrel{\mathrm{def}}{=} \frac{\mathbf{s}^{\mathrm{b}(l)} + \mathbf{s}^{\mathrm{r}(l)} + \mathbf{s}^{\mathrm{t}(l)} + \mathbf{b}^{(l)}}{1 + z_t^{(l-1)} + z_{t-1}^{(l)}} \tag{A7}$$

In this way, information is only passed upward and downward at boundaries, and the boundaries thus govern information flow between adjacent layers.

The vector $\mathbf{s}_t^{(l)}$ is split into state preactivations $\mathbf{u}_t^{(l)} \in \mathbb{R}^{4D_l}$ and scalar boundary preactivation $v_t^{(l)} \in \mathbb{R}$. Discrete boundary $z_t^{(l)}$ is computed from $v_t^{(l)}$ stochastically during training:

$$z_t^{\mathrm{train}(l)} \sim \mathrm{Bernoulli}\left(\mathrm{sigmoid}(v_t^{(l)})\right) \tag{A8}$$

and deterministically during evaluation:

$$z_t^{\mathrm{eval}(l)} = \begin{cases} 1 & v_t^{(l)} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{A9}$$

We additionally require that $z_t^{(0)} = 1$ (the input always "segments") and $z_t^{(L)} = 0$ (the top layer never segments). To enforce hierarchical segmentation behavior, we mask the boundaries by the boundaries at the layer below:

$$z_t^{(l)} \leftarrow z_t^{(l)} z_t^{(l-1)} \tag{A10}$$

Gradients through these discrete decisions are approximated using straight-through estimation (Hinton, 2012; Bengio et al., 2013; Courbariaux et al., 2016; Chung et al., 2017; Shain and Elsner, 2019; Eloff et al., 2019).

The forget gates $\mathbf{f}_t^{\mathrm{e}(l)}$, input gates $\mathbf{i}_t^{\mathrm{e}(l)}$, output gates $\mathbf{o}_t^{\mathrm{e}(l)}$, and cell proposal $\mathbf{g}_t^{\mathrm{e}(l)}$ are computed as follows:

$$\mathbf{f}_t^{\mathrm{e}(l)} \stackrel{\mathrm{def}}{=} \mathrm{sigmoid}\left(\mathbf{u}_t^{(l)}{}_{[1:D_l]}\right) \tag{A11}$$

$$\mathbf{i}_t^{\mathrm{e}(l)} \stackrel{\mathrm{def}}{=} \mathrm{sigmoid}\left(\mathbf{u}_t^{(l)}{}_{[D_l+1:2D_l]}\right) \tag{A12}$$

$$\mathbf{o}_t^{\mathrm{e}(l)} \stackrel{\mathrm{def}}{=} \mathrm{sigmoid}\left(\mathbf{u}_t^{(l)}{}_{[2D_l+1:3D_l]}\right) \tag{A13}$$

$$\mathbf{g}_t^{\mathrm{e}(l)} \stackrel{\mathrm{def}}{=} \tanh\left(\mathbf{u}_t^{(l)}{}_{[3D_l+1:4D_l]}\right) \tag{A14}$$

The cell state $\mathbf{c}_t^{\mathrm{e}(l)}$ is a weighted sum of three terms: a *flush* operation $\mathbf{c}_t^{\mathrm{f}(l)}$ that erases the cell memory, a standard LSTM *update* $\mathbf{c}^{\mathrm{u}(l)}$, and a *copy* operation $\mathbf{c}^{\mathrm{c}(l)}$ that copies the preceding cell state forward:

$$\mathbf{c}_t^{\mathrm{f}(l)} \stackrel{\text{def}}{=} \mathbf{i}_t^{\mathrm{e}(l)} \odot \mathbf{g}_t^{(\mathrm{e}l)} \tag{A15}$$

$$\mathbf{c}_t^{\mathrm{u}(l)} \stackrel{\text{def}}{=} \mathbf{f}_t^{\mathrm{e}(l)} \odot \mathbf{c}_{t-1}^{\mathrm{e}(l)} + \mathbf{i}_t^{\mathrm{e}(l)} \odot \mathbf{g}_t^{\mathrm{e}(l)} \tag{A16}$$

$$\mathbf{c}_t^{\mathrm{c}(l)} \stackrel{\text{def}}{=} \mathbf{c}_{t-1}^{\mathrm{e}(l)} \tag{A17}$$

These terms are weighted by the boundary decision such that a flush occurs when the preceding timestep finds a boundary, an update otherwise occurs when the layer below finds a boundary, and a copy occurs when neither layer finds a boundary:

$$\begin{aligned}\mathbf{c}_t^{\mathrm{e}(l)} \stackrel{\text{def}}{=} & z_{t-1}^{(l)} \mathbf{c}_t^{\mathrm{f}(l)} + \\ & \left(1 - z_{t-1}^{(l)}\right) z_t^{(l-1)} \mathbf{c}_t^{\mathrm{u}(l)} + \\ & \left(1 - z_{t-1}^{(l)}\right) \left(1 - z_t^{(l-1)}\right) \mathbf{c}_t^{\mathrm{c}(l)}\end{aligned} \tag{A18}$$

The hidden state $\mathbf{h}_t^{\mathrm{e}(l)}$ is computed as:

$$\mathbf{h}_t^{\mathrm{e}(l)} \stackrel{\text{def}}{=} \tanh\left(\mathbf{c}_t^{\mathrm{e}(l)}\right) \tag{A19}$$

The previous segment encoding $\mathbf{h}_t^{e(l)'}$ is updated following a boundary and copied forward otherwise:

$$\mathbf{h}_t^{\mathrm{e}(l)'} \stackrel{\text{def}}{=} z_{t-1}^{(l)} \mathbf{h}_{t-1}^{\mathrm{e}(l)} + \left(1 - z_{t-1}^{(l)}\right) \mathbf{h}_{t-1}^{\mathrm{e}(l)'} \tag{A20}$$

The current segment length $n_t^{(l)}$ is reset to 0 if $z_{t-1}^{(l)} = 1$, incremented if $z_t^{(l-1)} = 1$, and copied forward otherwise:

$$n_t^{(l)} \stackrel{\text{def}}{=} \left(1 - z_{t-1}^{(l)}\right) n_{t-1}^{(l)} + z_t^{(l-1)} \tag{A21}$$

B Comparison of Encoder Model to Chung et al. (2017)

Although our encoder model closely follows the definition in Chung et al. (2017), it differs in the following ways:

- The recurrent connection includes both the previous segment label and the current segment length in addition to the previous hidden state. We found this to be helpful during model development, and we hypothesize that this is because doing so removes the need for this information to be encoded by the model.
- We implement the case-wise reasoning of the segmentation decisions using multiplicative masking rather than logical selection. This is intended to boost signal into the boundary decisions.
- We enforce hierarchical segmentation behavior by multiplicatively masking the segmentation decision at layer l with the segmentation decision at layer $l-1$, thus preventing higher layers from segmenting where lower layers do not.
- We compute boundaries during training via Bernoulli sampling rather than rounding. We found this to substantially improve performance on the development set, and we hypothesize that sampling may improve the straight-through gradient estimates by ensuring that the segmentation decision is unbiased with respect to the underlying segmentation probability.
- We renormalize the preactivations $\mathbf{s}_t^{(l)}$ by the incoming boundary decisions (eq. A7). We found this to be helpful during model development, and we hypothesize that this is because it avoids fluctuation in the scale of preactivations as a function of the boundaries.
- We do not apply the Chung et al. (2017) technique of *slope annealing*, i.e. gradually increasing the steepness of the sigmoid activation function to reduce bias in the straight-through estimator. We did not find an appreciable benefit from slope annealing during development, and it had a tendency to produce training instability. Eliminating it also reduces experimenter degrees of freedom by removing design decisions about the annealing function.

C Decoder Definition

The decoder consists of two attentional seq2seq LSTMs with L layers each, one backward-directional (memory) and one forward directional (prediction). Given a backward window size B and a forward window size F, each backward decoder layer generates reconstructions $\mathbf{Y}_t^{\mathrm{B}(l)} \in \mathbb{R}^{B \times D_{l-1}}$ and each forward decoder layer generates predictions $\mathbf{Y}_t^{\mathrm{F}(l)} \in \mathbb{R}^{F \times D_{l-1}}$, corresponding respectively to the B preceding and F following segment labels of layer $l-1$ at time t. The initial decoder hidden and cell states — $\mathbf{h}_{t,0}^{\mathrm{dB}(l)}$ and $\mathbf{c}_{t,0}^{\mathrm{dB}(l)}$ for the

backward decoder and $\mathbf{h}_{t,0}^{\mathrm{dB}(l)}$ and $\mathbf{c}_{t,0}^{\mathrm{dF}(l)}$ for the forward decoder — are generated using multilayer feedforward transforms $f^{\mathrm{hB}(l)}$, $f^{\mathrm{cB}(l)}$, $f^{\mathrm{hF}(l)}$, and $f^{\mathrm{cF}(l)}$:

$$\mathbf{h}_{t,0}^{\mathrm{dB}(l)} \stackrel{\mathrm{def}}{=} f^{\mathrm{hB}(l)}\left(\mathbf{h}_t^{\mathrm{e}(l)}\right) \tag{A22}$$

$$\mathbf{c}_{t,0}^{\mathrm{dB}(l)} \stackrel{\mathrm{def}}{=} f^{\mathrm{cB}(l)}\left(\mathbf{h}_t^{\mathrm{e}(l)}\right) \tag{A23}$$

$$\mathbf{h}_{t,0}^{\mathrm{dF}(l)} \stackrel{\mathrm{def}}{=} f^{\mathrm{hF}(l)}\left(\mathbf{h}_t^{\mathrm{e}(l)}\right) \tag{A24}$$

$$\mathbf{c}_{t,0}^{\mathrm{dF}(l)} \stackrel{\mathrm{def}}{=} f^{\mathrm{cF}(l)}\left(\mathbf{h}_t^{\mathrm{e}(l)}\right) \tag{A25}$$

Decoder states are doubly time indexed by t, i, where t indexes the encoder timestamp (i.e. the input timestep at which decoding begins) and i indexes the decoder timestamp (i.e. progress through the B or F decoder frames). Given decoder states $\mathbf{h}_{t,i}^{\mathrm{dB}(l)}$, $\mathbf{h}_{t,i}^{\mathrm{dF}(l)}$, predictions $\mathbf{Y}_t^{\mathrm{B}(l)}{}_{[i]}$, $\mathbf{Y}_t^{\mathrm{F}(l)}{}_{[i]} \in \mathbb{R}^{D_{l-1}}$ are generated using multilayer feedforward transforms $f^{\mathrm{yB}(l)}$, $f^{\mathrm{yF}(l)}$:

$$\mathbf{Y}_t^{\mathrm{B}(l)}{}_{[i]} \stackrel{\mathrm{def}}{=} f^{\mathrm{yB}(l)}\left(\mathbf{h}_{t,i}^{\mathrm{dB}(l)}\right) \tag{A26}$$

$$\mathbf{Y}_t^{\mathrm{F}(l)}{}_{[i]} \stackrel{\mathrm{def}}{=} f^{\mathrm{yF}(l)}\left(\mathbf{h}_{t,i}^{\mathrm{dF}(l)}\right) \tag{A27}$$

The decoder takes as input a periodic positional encoding $\mathbf{e}_i$, generated following Vaswani et al. (2017). Non-final layers additionally take as an attention values the predictions from the layer above, i.e. $\mathbf{Y}_t^{\mathrm{B}(l+1)}$ (for backward reconstruction) and $\mathbf{Y}_t^{\mathrm{F}(l+1)}$ (for forward prediction) and compute a weighted sum of these values over time with attention weight vectors $\mathbf{a}_{t,i}^{\mathrm{B}(l)} \in (0,1)^B$ and $\mathbf{a}_{t,i}^{\mathrm{F}(l)} \in (0,1)^F$ to generate context vectors $\mathbf{w}_{t,i}^{\mathrm{B}(l)}$ and $\mathbf{w}_{t,i}^{\mathrm{F}(l)}$:

$$\mathbf{w}_{t,i}^{\mathrm{B}(l)} \stackrel{\mathrm{def}}{=} \mathbf{Y}_t^{\mathrm{B}(l+1)\top}\mathbf{a}_{t,i}^{\mathrm{B}(l)} \tag{A28}$$

$$\mathbf{w}_{t,i}^{\mathrm{F}(l)} \stackrel{\mathrm{def}}{=} \mathbf{Y}_t^{\mathrm{F}(l+1)\top}\mathbf{a}_{t,i}^{\mathrm{F}(l)} \tag{A29}$$

Attention weights $\mathbf{a}^{\mathrm{B}(l)}$ and $\mathbf{a}^{\mathrm{F}(l)}$ are computed using Gaussian kernel $k(i;\mu,\sigma^2)$:

$$k(i;\mu,\sigma^2) \stackrel{\mathrm{def}}{=} \exp\left(\frac{(i-\mu)^2}{\sigma^2}\right) \tag{A30}$$

Kernel k is applied to decoder time, with concentration $\sigma^{\mathrm{B}(l)}, \sigma^{\mathrm{F}(l)} = 0.25$ and with location $\mu_{t,i}^{\mathrm{B}(l)}$ $\mu_{t,i}^{\mathrm{F}(l)} \in \mathbb{R}_+$ computed by transforming the previous decoder state using a feedforward transform $f^{\mathrm{qB}(l)}$, $f^{\mathrm{qF}(l)}$ and adding the result to the previous attention location:

$$\mu_{t,i}^{\mathrm{B}(l)} \stackrel{\mathrm{def}}{=} \mathrm{abs}\left(f^{\mathrm{qB}(l)}\left(\mathbf{h}_{t,i-1}^{\mathrm{dB}(l)}\right)\right) + \mu_{t,i-1}^{\mathrm{B}(l)} \tag{A31}$$

$$\mu_{t,i}^{\mathrm{F}(l)} \stackrel{\mathrm{def}}{=} \mathrm{abs}\left(f^{\mathrm{qF}(l)}\left(\mathbf{h}_{t,i-1}^{\mathrm{dF}(l)}\right)\right) + \mu_{t,i-1}^{\mathrm{F}(l)} \tag{A32}$$

where $\mu_{t,0}^{\mathrm{B}(l)} = 1$. Unit-normalized attention vectors are computed from timestamp vectors $\mathbf{t}^{\mathrm{B}} \stackrel{\mathrm{def}}{=} (1,\ldots,B)^\top$ and $\mathbf{t}^{\mathrm{F}} \stackrel{\mathrm{def}}{=} (1,\ldots,F)^\top$ as:

$$\mathbf{a}_{t,i}^{\mathrm{B}(l)} \stackrel{\mathrm{def}}{=} \frac{k\left(\mathbf{t}^{\mathrm{B}};\sigma^{\mathrm{B}(l)},\mu_{t,i}^{\mathrm{B}(l)}\right)}{\sum_{j=1}^{B} k\left(\mathbf{t}_{[j]}^{\mathrm{B}};\sigma^{\mathrm{B}(l)},\mu_{t,j}^{\mathrm{B}(l)}\right)} \tag{A33}$$

$$\mathbf{a}_{t,i}^{\mathrm{F}(l)} \stackrel{\mathrm{def}}{=} \frac{k\left(\mathbf{t}^{\mathrm{F}};\sigma^{\mathrm{F}(l)},\mu_{t,i}^{\mathrm{F}(l)}\right)}{\sum_{j=1}^{B} k\left(\mathbf{t}_{[j]}^{\mathrm{F}};\sigma^{\mathrm{F}(l)},\mu_{t,j}^{\mathrm{F}(l)}\right)} \tag{A34}$$

The attention weights are thus constrained to march monotonically in time from t into the decoded past or future predicted segment labels from the layer above. Using fixed concentration 0.25 yields an effective kernel width $[-2\sigma, 2\sigma]$ of one timestep, ensuring that the bulk of the attention kernel either falls on a single segment label or straddles two consecutive segment labels and preventing the decoder from spreading its attention over many higher-level segments. This design encourages one-to-many temporal alignment between decoded segment labels and decoded inputs, while allowing the decoder to determine how long to attend to a predicted segment label before moving on to the next one. At the final (top) layer, no top-down predictions are available, so the context vectors are omitted (or, equivalently, set to $\mathbf{0}$).

The inputs to the decoder $\mathbf{x}_{t,i}^{\mathrm{dB}(l)}$, $\mathbf{x}_{t,i}^{\mathrm{dF}(l)}$ are constructed as the vertical concatenation of $\mathbf{e}$, $\mathbf{w}$, and the previously generated decoder output, and a standard LSTM state update is applied:

$$\mathbf{x}_{t,i}^{\mathrm{dB}(l)} \stackrel{\mathrm{def}}{=} \begin{pmatrix} \mathbf{e}_i \\ \mathbf{w}_{t,i}^{\mathrm{B}(l)} \\ \mathbf{Y}_t^{\mathrm{B}(l)}{}_{[i-1]} \end{pmatrix} \tag{A35}$$

$$\mathbf{x}_{t,i}^{\mathrm{dF}(l)} \stackrel{\mathrm{def}}{=} \begin{pmatrix} \mathbf{e}_i \\ \mathbf{w}_{t,i}^{\mathrm{F}(l)} \\ \mathbf{Y}_t^{\mathrm{F}(l)}{}_{[i-1]} \end{pmatrix} \tag{A36}$$

$$\mathbf{h}_{t,i}^{\mathrm{dB}(l)}, \mathbf{c}_{t,i}^{\mathrm{dB}(l)} \stackrel{\mathrm{def}}{=} \mathrm{LSTM}\left(\mathbf{x}_{t,i}^{\mathrm{dB}(l)}, \mathbf{h}_{t,i-1}^{\mathrm{dB}(l)}\right), i > 0 \tag{A37}$$

$$\mathbf{h}_{t,i}^{\mathrm{dF}(l)}, \mathbf{c}_{t,i}^{\mathrm{dF}(l)} \stackrel{\mathrm{def}}{=} \mathrm{LSTM}\left(\mathbf{x}_{t,i}^{\mathrm{dF}(l)}, \mathbf{h}_{t,i-1}^{\mathrm{dF}(l)}\right), i > 0 \tag{A38}$$

The decoder is only applied to elements of $f^{\mathrm{f}(l)}(1, T)$ (i.e. only to frames where layer l segments) and only decodes the last B elements of $f^{\mathrm{f}(l-1)}(1, t)$ and the first F elements of $f^{\mathrm{f}(l-1)}(t+1, T)$; that is, it decodes only the B preceding segment labels and F following segment labels from layer $l-1$, ignoring labels at non-boundaries. Therefore, like encoding, decoding is also multi-scale, taking place at the timescale of the encoder representations.

D Objective

Each decoder layer contributes two terms to the objective, a forward objective and a backward objective. Layer 1 decodes the data and uses a squared error loss:

$$f^{\mathcal{L}(1)}(x, y) \stackrel{\text{def}}{=} ||x - y||_2^2 \tag{A39}$$

Layers $2, \ldots, L$ decode the representations from the layer below, which are tanh-activated and thus constrained to the interval $(-1, 1)$. Encoder features $\mathbf{h}_t^{\mathrm{e}(l)}$ are deterministically cast into bitwise feature probabilities $\mathbf{p}_t^{\mathrm{e}(l)}$ and decoded using sigmoid cross-entropy loss:

$$\mathbf{p}_t^{\mathrm{e}(l)} \stackrel{\text{def}}{=} (\mathbf{h}_t^{\mathrm{e}(l)} + 1)/2 \tag{A40}$$

$$f^{\mathcal{L}(l)}(x, y) \stackrel{\text{def}}{=} \text{sigmoid-xent}(x, y), 1 < l \leq L \tag{A41}$$

Let $T^{(l)\prime}$ denote the number of segment boundaries in layer l. Let $\mathbf{Z}^{\mathrm{B}(l)}_{t,i,d}$, $\mathbf{Z}^{\mathrm{F}(l)}_{t,i,d}$, $\hat{\mathbf{Z}}^{\mathrm{B}(l)}_{t,i,d}$, and $\hat{\mathbf{Z}}^{\mathrm{F}(l)}_{t,i,d}$ respectively denote the backward and forward targets and model predictions at encoder time t, decoder time i, and dimension d, defined as follows:

$$\mathbf{Z}^{\mathrm{B}(l)}_{t,i,d} \stackrel{\text{def}}{=} \text{rev}\left(f^{\mathrm{f}(l-1)}(1, t)\right)_{[i,d]} \tag{A42}$$

$$\mathbf{Z}^{\mathrm{F}(l)}_{t,i,d} \stackrel{\text{def}}{=} f^{\mathrm{f}(l-1)}(t+1, T)_{[i,d]} \tag{A43}$$

$$\hat{\mathbf{Z}}^{\mathrm{B}(l)}_{t,i,d} \stackrel{\text{def}}{=} \mathbf{Y}^{\mathrm{B}(l)}_{t\ [i,d]} \tag{A44}$$

$$\hat{\mathbf{Z}}^{\mathrm{F}(l)}_{t,i,d} \stackrel{\text{def}}{=} \mathbf{Y}^{\mathrm{F}(l)}_{t\ [i,d]} \tag{A45}$$

The backward and forward loss components $\mathcal{L}^{\mathrm{B}(l)}$ and $\mathcal{L}^{\mathrm{F}(l)}$ are computed as:

$$\mathcal{L}^{\mathrm{B}(l)} \stackrel{\text{def}}{=} \frac{\sum_{t=1}^{T^{(l)\prime}} \sum_{i=1}^{B} \sum_{d=1}^{D_{l-1}} f^{\mathcal{L}(l)}\left(\mathbf{Z}^{\mathrm{B}(l)}_{t,i,d}, \hat{\mathbf{Z}}^{\mathrm{B}(l)}_{t,i,d}\right)}{T^{(l)\prime} B D_{l-1}} \tag{A46}$$

$$\mathcal{L}^{\mathrm{F}(l)} \stackrel{\text{def}}{=} \frac{\sum_{t=1}^{T^{(l)\prime}} \sum_{i=1}^{F} \sum_{d=1}^{D_{l-1}} f^{\mathcal{L}(l)}\left(\mathbf{Z}^{\mathrm{F}(l)}_{t,i,d}, \hat{\mathbf{Z}}^{\mathrm{F}(l)}_{t,i,d}\right)}{T^{(l)\prime} F D_{l-1}} \tag{A47}$$

The overall loss $\mathcal{L}$ is:

$$\mathcal{L} \stackrel{\text{def}}{=} \sum_{l=1}^{L} \mathcal{L}^{\mathrm{B}(l)} + \mathcal{L}^{\mathrm{F}(l)} \tag{A48}$$

E Implementation Details

We apply the following implementation decisions in this study:

- $D_l = 128$ for $1 \leq l \leq L$
- One hidden layer of 128 units for all feedforward transforms
- Positional encoding dimensionality of 128
- Exponential linear unit (elu) activations for all internal feedforward layers (Clevert et al., 2015)
- Glorot uniform initialization for bottom-up, top-down, and feedforward encoder and decoder weight matrices (Glorot and Bengio, 2010)
- Orthogonal initialization for recurrent weight matrices (Saxe et al., 2013)
- Adam optimizer (Kingma and Ba, 2014) with learning rate 0.001, a minibatch size of 8, and default TensorFlow parameters.
- **Probing classifier implementation**
  - Logistic regression using `scikit-learn` (Pedregosa et al., 2011)
  - Phoneme prediction is multinomial, feature prediction is binary
  - Minority feature class is always coded as positive
  - 2-fold cross-validation
  - L2 $\lambda = 1$
  - 100 LBFGS iterations (Zhu et al., 1997) per fold

F Data Preprocessing

We convert the audio recordings into sequences of 50-dimensional cochleagrams (Brown and Cooke, 1994; McDermott and Simoncelli, 2011), each representing 10ms of audio data. Although this differs from the standard automatic speech recognition pipeline based on Mel frequency cepstral coefficients (Mermelstein, 1976), it is motivated for our

study because the model is unsupervised. Since we wish to test theories about cognition by extracting features from the acoustic stream without supervision, it is critical not only that the speech representation contain features that support identification of linguistic units, but that the representation emphasize those features in a plausibly similar manner to that of the human auditory system. Cochleagrams support this goal by incorporating more recent insights about human auditory perception (McDermott and Simoncelli, 2011). Our implementation uses the `pycochleagram` library https://github.com/mcdermottLab/pycochleagram.

We L2 normalize the cochleagrams in order to encourage the decoder to focus on the spectral power envelope rather than absolute variation in loudness, since the former plausibly contains more linguistic signal. This procedure is supported by evidence of loudness constancy in human auditory perception, suggesting that similar kinds of normalization may take place in the brain (Zahorik and Wightman, 2001). We additionally z-transform the normalized cochleagrams over time within each audio file, since this proved beneficial during model development.

The source audio files contain many non-speech regions that are not of direct relevance for this study. We use the voice activity detection (VAD) intervals provided with the Zerospeech 2015 challenge data to remove these regions as a preprocess, and we force boundaries at the ends of VAD intervals. This greatly speeds training by removing irrelevant data, and it aligns with neuroscientific evidence of a prelinguistic capacity to detect human voices (Belin et al., 2000; Fecteau et al., 2005; Blasi et al., 2011; Pernet et al., 2015).

G Regression Model Design and Results

We use linear regression to test the relationship between performance and memory pressures, prediction pressures, and multiscale encoding. To do so, we combine raw boundary, phoneme classification, and feature classification metrics, along with deltas in these metrics over baselines U and X, into a single vector of performance statistics, each of which measures one aspect of the contribution of these dimensions to phoneme learning in our unsupervised models. To improve normality of performance metrics which are bounded on the interval $[0, 1]$, as well as comparability of performance across metrics, we first (1) cast the metrics onto the interval $[-1, 1]$, (2) apply Fisher's Z transformation (i.e. arctanh), and (3) Z-score the transformed vectors within each metric type.

Predictor	β	t	p
Intercept	-1.22	-7.73	**3.89e-14*****
Memory	0.247	2.75	**0.006****
Prediction	0.959	9.86	**2.0e-16*****
Multiscale	0.305	4.10	**4.58e-5*****
Comparison=Full	0.037	0.453	0.651
Comparison=BaselineX	-0.064	-0.709	0.479
Metric=Phoneme	0.021	0.240	0.810
Metric=Feature	0.022	0.250	0.803

Table A1: Linear regression results

We use binary coding for our predictors of interest: presence/absence of memory pressures ($B > 0$), presence/absence of prediction pressures ($F > 0$), and presence/absence of multiscale segmental encoding ($L > 2$). We also include categorical controls for comparison type (full, full - baseline U, full - baseline X) and metric type (boundary, phoneme, feature). Results, shown in Table A1, support a contribution of all three critical variables to phoneme acquisition.