COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13427

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13427

Computational Modeling of the Segmentation of Sentence Stimuli From an Infant Word-Finding Study

Daniel Swingley,[a] Robin Algayres[b]

[a]*Department of Psychology, University of Pennsylvania*
[b]*ENS, Paris Sciences et Lettres*

Received 17 November 2023; received in revised form 22 February 2024; accepted 24 February 2024

Abstract

Computational models of infant word-finding typically operate over transcriptions of infant-directed speech corpora. It is now possible to test models of word segmentation on speech materials, rather than transcriptions of speech. We propose that such modeling efforts be conducted over the speech of the experimental stimuli used in studies measuring infants' capacity for learning from spoken sentences. Correspondence with infant outcomes in such experiments is an appropriate benchmark for models of infants. We demonstrate such an analysis by applying the DP-Parser model of Algayres and colleagues to auditory stimuli used in infant psycholinguistic experiments by Pelucchi and colleagues. The DP-Parser model takes speech as input, and creates multiple overlapping embeddings from each utterance. Prospective words are identified as clusters of similar embedded segments. This allows segmentation of each utterance into possible words, using a dynamic programming method that maximizes the frequency of constituent segments. We show that DP-Parse mimics American English learners' performance in extracting words from Italian sentences, favoring the segmentation of words with high syllabic transitional probability. This kind of computational analysis over actual stimuli from infant experiments may be helpful in tuning future models to match human performance.

Keywords: Infant language; Word recognition; Speech segmentation; Computational modeling; Zero-resource speech; Developmental psycholinguistics

Correspondence should be sent to Daniel Swingley, Department of Psychology, University of Pennsylvania, 425 S. University Ave, Philadelphia, PA 19104, USA. E-mail: swingley@psych.upenn.edu

1. Introduction

The ideal theory of language acquisition would be one that allowed us to make predictions about the course of development in individual children learning any language or languages. By knowing details of the child's language environment, and by measuring aspects of the child's disposition and ability, we might foretell characteristics of their language performance, and have some purchase on how changes in the language environment might alter this performance. The ideal theory would also be *mechanistic*, offering not only prediction but explanation.

Considering the earliest phase of language development, it has been difficult to approach a theory of this sort. In our view, the three largest impediments have been the inadequacy of computational methods for handling the speech signal, the lack of a quantitative theory of real-world concept formation, and the absence of closely annotated corpora. In principle, if we knew how infants interpret the speech signal and how they draw generalizations over other aspects of their perceptual experience, we could take enormous strides in "reverse engineering" the cognitive process of language acquisition (Dupoux, 2018).

Here, we attempt to justify a "reverse engineering" approach, and then provide an example of a new kind of computational modeling effort. In essence, we try to mimic elements of the cognitive process of infants participating in a speech segmentation experiment, by presenting a learning model with the actual speech materials from that experiment, and probing the outputs of the model to evaluate its alignment with the representations presumed to underlie infants' behavior in the experiments. The computational model is that of Algayres et al. (2022), and the infant experiments are two studies reported in Pelucchi, Hay, and Saffran (2009a, 2009b). In these experiments, 8-month-old infants heard fluent sentences of an unfamiliar language (Italian) and extracted words from these sentences. Infants favored words whose syllables were statistically cohesive.

These studies are appropriate targets of modeling for two reasons. First, they used naturally produced full sentences of a real language, unlike the vast majority of experiments measuring infant "statistical learning," which render speech using synthesis or mechanical concatenation of syllables of recorded human speech. Highly artificial materials like these usually present learners with token-identical types (a syllable is exactly the same every time it happens) and with unnaturally uniform rhythms (which may produce a stiff temporal entrainment alien to normal language processing). Real sentences reflect better the challenge to computational modeling that our work here addresses. Learning in context demands that infants discover where the relevant units are, and which instances should be treated as the same for generalization (e.g., Yates et al., 2022). Second, Pelucchi et al. went beyond testing sensitivity to word frequency. As described below, the target and distracter words in the stimuli occurred equally often, differing only in the relative conditional probabilities of the words' two syllables. This allows for a more stringent test of the similarity between infants and the model.

Our work here is, fundamentally, a demonstration project, joining other work that aims to harness developments in computational speech technology (e.g., Elsner, Goldwater, & Eisenstein, 2012; Matusevych, Schatz, Kamper, Feldman, & Goldwater, 2023; Räsänen & Rasilo, 2015; Roy & Pentland, 2002) with precursors in the machine learning literature (e.g.,

Harwath, Torralba, & Glass, 2016). We view our approach as contrasting with the direction that language acquisition research usually takes. The scientific study of language acquisition usually starts with the mature language being acquired, and works its way backwards. Researchers take the phonemes, words, morphological elements, and syntactic regularities characterized by linguistic analysis, and, at each level, attempt to trace the learning of these components in children. This has been a productive strategy for decades. For example, by defining the consonants and vowels of the infant's home language, we can test when infants show signs of favoring these sounds over the sound contrasts of other languages, which they start to do within the first year (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Werker & Tees, 1984). By counting how often sequences of these discrete units appear together, we can test when infants begin to react differently to frequent and infrequent sequences, showing infants' absorption of native-language sequential probabilities (Friederici & Wessels, 1993; Mattys, Jusczyk, Luce, & Morgan, 1999). And by knowing the stock of common words in typical children's language environments, we can examine infants' preference for these forms over rarer ones (Hallé & de Boysson-Bardies, 1994). Broadly speaking, the targets-focused approach has revealed many surprisingly precocious developments, pulling developmental achievement milestones to earlier points than might be predicted on the basis of children's spoken language.

However, this research approach has risks. In our view, starting from the target and working backward (a) tends to result in the documentation of milestones rather than characterization of the gradual mechanics of developmental processes; and (b) tends to favor a serial approach in which infants are assumed to acquire linguistic levels in a logical order: first, speech sounds, then word-forms, then word meanings, then morphosyntactic regularities. Two results have persuaded us that this is a problem. First, it appears that infants learn words during the same period that they are first beginning to show phonetic adaptation to their native language (Bergelson & Swingley, 2012). This means that there is a period in which the lexicon is present, but words in the lexicon are probably not encoded in terms of language-specific phonological units—or, to follow an interpretation long present in the literature, perhaps not in phonological units at all (Beckman & Edwards, 2000; Jusczyk, 1993; Vihman & Croft, 2007; Werker & Curtin, 2005; see also Feldman, Goldwater, Dupoux, & Schatz, 2021). This, in turn, implies that early word-finding does not work in the manner proposed by virtually every quantitative model of infant word-form learning in the psycholinguistic literature. The models start from language-specific phonetic categories (first demonstrated empirically between 6 and 12 months), sometimes aggregated into language-specific syllables or divided into phonological feature bundles, and compute statistics over these, implementing heuristics which, together with some phonetic boundary cues, help specify word boundaries (Bernard et al., 2020). If these characterizations of representations are inaccurate, then the models are wrong.

Second, even if we were to assume that word-form learning begins after children have made good progress on learning their language's consonants and vowels, the speech directed to infants does not reliably contain those sounds articulated such that even adults can identify them. For example, Swingley (2021) found that native English-speaking adults listening to infant-directed speech could identify only about half of word-onset consonants more than half of the time, and only about a third of word-final consonants (by tokens). Consonants were

presented in vowel-consonant-vowel contexts (or, in a follow-up, in VCV contexts with the rest of the sentence provided in a low-pass-filtered form), with no demand for speed. If adult native speakers cannot reliably identify the sounds of infant-directed speech, infants cannot either. Therefore, we cannot simply assume that infants can recover from infant-directed speech a sequence of canonically categorized consonants and vowels to be used as inputs to a statistical learning process as it is usually described in the psycholinguistic modeling literature.

Furthermore, even to the extent that phonological strings can be identified, speech is also characterized by many-to-many mappings between these phonological strings and the words (lemmas) intended by the speaker. This is a problem, because many words are phonologically similar to one another. As a result, words often share reduced pronunciations with other words' canonical and reduced forms: for example, the three words *heel, he'll*, and *hill* all share commonly attested pronunciations in American English. These three types are not unusual—such variation-driven entanglement is the predominant case (2/3 of types, 85% of tokens; Beech & Swingley, 2023). Even if children could identify speech sounds in context the same way native-speaker corpus transcribers do, children would need additional principles to sort out phonology and the lexicon.

These considerations indicate that our attempts to account for vocabulary development and other components of the infant language learning process cannot be accurate if they start from presumption that speech is made of sequences of identifiable phones delivered to a sequential statistical learning algorithm. While freely acknowledging our own prior modeling efforts of this sort (Swingley, 1999, 2005), it is time to explore something new.

The best alternative would be to start from the acoustic signal (and, eventually, other features of the language environment), and attempt to model the developmental process from the infant's perspective, rather than from starting with the target and working backwards. We are not the first to suggest this strategy (see the examples cited above), but this remains a minority approach in the infancy literature, and has yet to be well integrated with psycholinguistic approaches outside the computer lab, particularly if we consider work that starts from the speech signal itself.

The primary obstacle to a phonetic reverse-engineering strategy has always been that it is extremely difficult to model the naive human response to the speech signal. It is hard to convert speech recordings into a representation ("embedding") where the mathematical similarity between embeddings matches the perceptual similarity between the sounds that the embeddings represent. The success of modern speech models like the ones that underlie commercial systems like Alexa and Siri might seem to indicate that the embeddings problem has been solved, but these models are trained on massive amounts of labeled language data. They know the words and grammatical probabilities of the language already, and are taught to link phonetic patterns to these structures. Their successes are tightly bound to this omniscient training, and as a result, they do not present good models of human development.

However, significant progress has been made in the past 7–8 years or so on speech technologies intended to learn speech embeddings without significant top-down training on the target language. This research is primarily motivated by an interest in creating speech tools for languages for which there has been relatively little investment: minimal human

annotation, modest corpus resources, and small datasets. Without top-down information in quantity, the learning task is more like the infant's: to start from a generic speech embedding (trained, for the computer; innate, for the infant), and use phonetic regularities present in speech audio data to adapt to the target language. These "zero-resource" speech technologies are improving rapidly, and as a result, they are becoming more capable as models of human language learning (Dunbar et al., 2021).

One of the challenges of any modeling approach is in evaluation. A system may solve a task successfully without doing it in a humanlike way. In the case of infant language learning, this means that it is crucial to test a model according to what is known about infants' abilities, as revealed in their behavior. Given that infants' behavior in natural settings does not fully reveal their knowledge, researchers use observational experiments in controlled environments to estimate infants' knowledge and abilities. To the extent that a computational system mimics the behavior of children when presented with the same material, it suggests that the system could be useful for estimating children's knowledge state.

Here, we adopt this approach by applying the DP-Parse model of Algayres et al. (2022) to experimental stimuli previously used to test infants' learning from speech. The model is trained on samples of speech without being provided explicit phonetic categories, as described below, and, therefore, is a member of a class of self-supervised models that are suitable for mimicking aspects of the infant language acquisition process. The experimental stimuli came from observational studies of infants: we know what infants heard, and we know how they responded. Our goal, then, is to determine whether the model and the infants develop similar interpretations of these phonetic materials.

In the laboratory studies, American-English learning 8-month-olds listened to short, scripted passages of fluent Italian speech, and then lists of isolated words (one word type per trial). The passages contained many words, but were constructed to repeat four words six times each: *bici, casa, fuga*, and *melo*. In general, and in control experiments given in Pelucchi et al. (2009a), infants were expected to listen longer to *bici, bici, bici*…if they had just heard a passage with several *bici*s in it than if they had heard a passage featuring an alternative word in place of *bici*. The experimental studies asked: what makes a word a coherent unit for an 8-month-old, leading to this preference? The authors contrasted two notions: (a) *bici* is a word if it is heard frequently; or (b) *bici* is a word if its parts, *bi* and *ci*, tend to occur together; that is, *bi* is always followed by *ci* and *ci* is always preceded by *bi*.

These hypotheses were differentiated in the experiment by strategically inserting into the passages some words that contained a syllable from the target words, but without the partner syllable. Thus, for example, in Pelucchi, Hay, and Saffran (2009b), all four of these words occurred equally often (six times per iteration of the passage), but for two words, such as *bici* and *casa*, the perfect co-occurrence probabilities of the component syllables were broken by inserting 12 instances of *bi* and *ca* into the sentences as parts of other words. Because those 12 instances were not immediately before *ci* and *sa*, over the course of the passage, they disrupted the statistical coherence of *bici* and *casa*. If frequency of a bisyllable alone were what led infants in the baseline study to treat the bisyllable as familiar, the introduction of *bi* and *ca* should not have affected detection of the bisyllables; indeed, it might even facilitate their detection, perhaps by drawing attention to them, or reinvigorating them in memory. On

the other hand, if *bici* and *casa* were segmented as units because of their statistical coherence, infants should prefer words whose syllables occurred only within those words. This is what the authors found, both in the first study and in the second (Pelucchi et al., 2009a). In the latter paper's critical experiment, the syllables that disrupted the coherence of two of the targets were the *second* syllable of each word (e.g., *ci* and *sa*), which were made abundant in the passage. Infants again preferred words with unique and co-occurring syllables over equally frequent words with second-syllable "decoys."

The fact that the decoy syllables changed the availability of the target words constrains the set of explanations we might have about infants' extraction of information from speech in an unfamiliar language. The typical characterization of the results is that infants must break the speech signal into syllables or phones, and recognize various tokens of that unit as fitting into categories (e.g., each *bi* is recognized and counted as such). Given these category labels, infants compute their frequencies and co-occurrence frequencies, and either favor segments that are made from high-probability subsequences, or insert boundaries between segments of low probability.

There are other ways we might think about these cognitive processes. For example, infants might store in memory any sequence that is frequent, but break down the bisyllables with lower transitional probabilities because of dilution of the memory representation of the bisyllable given the frequent presentation of one of its components out of the bisyllable context. This kind of process would not require explicit computation of transitional probabilities. Models with this property include the text-based computational models of Perruchet and Vinter (1998) and Cabiddu, Bott, Jones, and Gambi (2023).

The infant experiments' methods do not permit sentence by sentence analysis—the sentences are presented during familiarization, and the experimental results come from infants' responses to isolated-word lists presented afterward. Because in the experiments the only stimulus difference between conditions concerns the familiarization, we assume that the right focus of modeling efforts is the set of familiarization sentences. Something about the high transitional probability (HTP) familiarization passages induces a preference for the HTP sequences later, which is not present (or less present) for the low transitional probability (LTP) passages.

We used DP-Parse to model infants' cognitive processes (Algayres et al., 2022). DP-Parse was inspired by the DP-Unigram model of Goldwater, Griffiths, and Johnson (2009), a nonparametric Bayesian model that uses a Dirichlet process to parse text data. DP-Unigram is based on a simple prior: a frequent sequence of letters is more likely to be a word than a rare one. Therefore, parses of sentences that consist of more frequent intervals are preferred over parses that contain less frequent intervals. In text data, the frequencies of letter sequences can be computed by counting the number of occurrences. In speech data, two new difficulties emerge. First, there are no explicit segmentation points. To deal with this, DP-Parse introduces hypothetical segmentation points every 40 ms (roughly half the duration of a phoneme).[1] Second, every speech segment is a unique piece of signal, which means that frequencies cannot be computed by counting occurrences. Algayres et al. proposed to estimate frequencies with a method based on speech sequence embeddings (SSE) and density estimation.[2]

An SSE is a fixed-size vector representation of a portion of speech. Like any cognitive representation, an SSE is effective when the mathematical differences among embeddings reflect the true differences in the space the representation encodes. As an analogy, we may consider geographic maps, which function well as representations because the distances and angles between objects on the map reflect the true distances and angles between objects in the world. In the case of speech, a good embedding space is one that places together speech segments that a human listener would consider similar. This is hard to achieve because the speech samples that adult native speakers of a language consider similar to another (or interchangeable with one another) are often not acoustically similar, on many simple or intuitive definitions of "similar." This may be true of a vowel spoken by a man or a woman, a consonant in one environment or another, or a word articulated either clearly and deliberately, or quickly and offhandedly.

Because the embedding space for speech is hard to get right, one approach is to train up the embeddings using millions of labeled examples. However, such a model would not be appropriate as a stand-in for human learning, because infants do not receive labeled speech data. An unsupervised system (or, more accurately for our purposes here, a self-supervised system) needs to train itself without it.

One way to do this is to employ the method of Algayres, Nabli, Sagot, and Dupoux (2023), who used the contrastive approach for acoustic word embeddings initiated by Livescu and colleagues to train a classifier without labeled training data (Kamper, Jansen, & Goldwater, 2016; Settle & Livescu, 2016). Here, we describe briefly how it works. The method starts by pulling intervals out of a speech corpus, and embedding them into a Wav2vec2 representation (Baevski, Zhou, Mohamed, & Auli, 2020). Wav2vec2 is a neural network trained on 1000 h of read English speech, a quantity that reasonably approximates the cumulative amount of speech heard by a North-American 8-month-old (Bergelson et al., 2019). This training was self-supervised, meaning that it did not rely on phonetic labels or any other top-down linguistic labels. We begin with Wav2vec2 because its speech embeddings are better at discriminating phonemes than simpler frequency-based speech representations, such as spectrograms or mel frequency cepstral coefficients (Hallap, Dupoux, & Dunbar, 2023).

The SSE model from Algayres et al. (2022) is a neural network trained on top of a frozen Wav2vec2 (i.e., Wav2vec2 parameters were kept unchanged). The speech intervals from the corpus are embedded with the Wav2vec2 representation after being distorted by manipulating their duration or pitch characteristics, creating acoustically new versions of each interval. The system is then trained to classify these acoustic variants as the same, while classifying speech intervals not deriving from the same instances as different. This training, called Noise Contrastive Estimation (Mnih & Teh, 2012), forces the neural network to create similar embeddings for pairs of speech sequences that sound similar. By doing so, the SSE model maximizes the phonetic information present in the input speech sequence, but minimizes irrelevant features such as those caused by speaker-specific vocal tract attributes. If the distortions that are used in training resemble the variability encountered during testing, this method provides a way to immunize against variability without top-down "cheating." In practice, the result is that two pronunciations of the same word are encoded similarly, and two different words are embedded farther away from each other.

The SSE model from Algayres (2022) could also be trained on simpler embeddings, such as Mel-filter banks or MFCCs, but the authors have shown the resulting SSEs have much lower word-level discriminative power (Algayres et al., 2022). Even though neural networks generally require a lot of training data, the SSE model from Algayres et al. has a small number of trainable parameters (the parameters of the Wav2vec2 model being frozen during training) and can be trained to reasonable performance with only a few spoken utterances, here less than 1 min of audio. Indeed, the SSE model is quite light in parameters, considering the space of similar models, as it is composed of only a one-dimensional convolution layer (512 channels, kernel size 4, and stride 1) and one transformer layer with embedding dimension 512 followed by a max pooling across time.

To achieve better performance, it is possible to incorporate some textual supervision in the training objective. Instead of creating pairs of speech chunks with manual distortions, as in the previous paragraph, pairs that have the same phonetic transcriptions can be created by leveraging time-aligned phonetic transcriptions. By training the same neural networks on those perfect pairs, the discriminative power of the resulting SSE model is increased. Our implementation of this kind of supervision involved contrastive training fed by selection of pairs of intervals that included the same string of phonological labels, but no internally marked gold-standard phone boundaries. This supervision steps away from the principle of never providing top-down teaching, and might be thought of as implementing a learner who adapts more quickly to a novel talker than our self-supervised model can. Still, weak supervision alters only the embeddings, and does not directly contribute to the segmentation task. In the present paper, we refer to this kind of added training as "weakly supervised," while the setting that does not use text annotations is referred to as "self-supervised." We focus primarily on the self-supervised model, but conclude with a comparison of the weakly supervised and self-supervised models.

This method of speech-sequence embedding, either the semi-supervised or weakly supervised one, is then applied to the sentences of the corpus. For each of the experimental passages, hypothetical word boundaries are placed every 40 ms. Word candidates are required to start and end at these 40-ms block boundaries and to be shorter than the maximal word duration (set here to 800 ms). The sequences, therefore, included all 40 ms intervals starting at {0, 40, 80,…ms}; all 80 ms intervals starting at {0, 40, 80,…ms}; and so on. The DP-Parse process as illustrated in Fig. 1 is composed of five steps. The first is to embed all of these speech intervals with the SSE model.

The second step is to obtain the frequencies of all speech intervals with respect to a set of already segmented words. At initialization, this set is empty, and so is temporarily filled with the intervals from a random segmentation of the corpus. This set of intervals is referred to as an "instance-based lexicon." It is not a proper lexicon, as it does not contain lemmas or types, but rather a collection of token speech intervals that is used to estimate the frequencies of elements encountered in new sentences. Of course, because each segment is represented with a unique high-dimensional vector, each literally has a frequency of zero (or one, if that segment already belongs to the lexicon), so we need a stand-in for frequency to apply to the parsing process. This was done using the Parzen–Rosenblatt window method (Parzen, 1962), as follows.

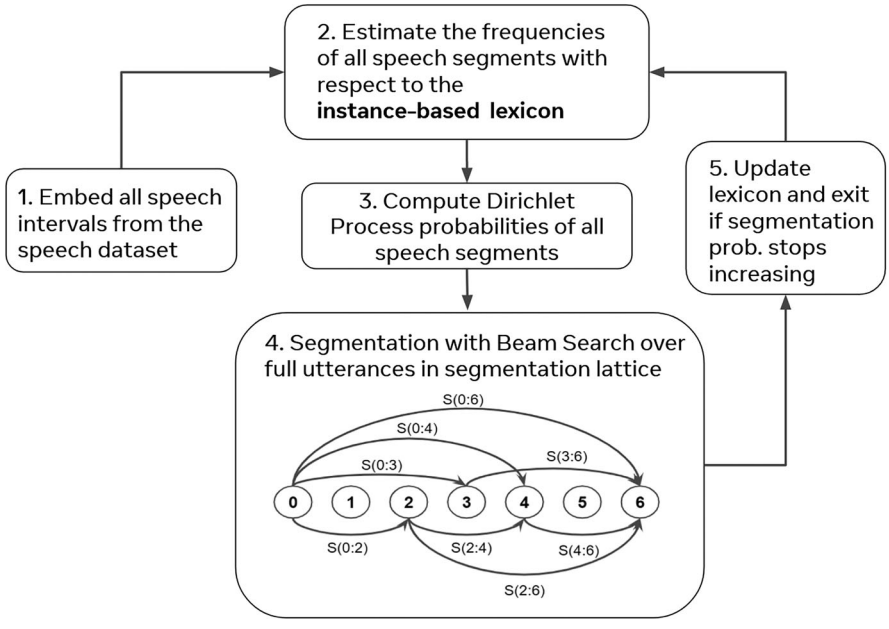

Fig. 1. Figure illustrating the DP-Parse process.

For each of the segments, the distance between each SSE and all of the others was computed, allowing grouping of vectors into sets using a k-NN (k nearest-neighbors) function. Then, for each vector, a Gaussian centered on that vector was fitted to the k-NN set, and the sum of the probability density function (pdf) of this Gaussian was computed. This pdf computation stood in for the frequency counting that could be done with discrete representations; in effect, it counted how many other vectors (speech chunks) were similar to the one being examined, weighting the counts by proximity. The only free parameter was the variance of the Gaussian, which was set to be the same for all estimated frequencies in the dataset. Algayres et al. (2022) introduced a simple prior for this purpose: that half of the speech segments in the collection of speech segments have a unique phonetic transcription. Therefore, the variance of the Gaussian was fitted so that half of the estimated frequencies were equal to 1. In Algayres, Zaiem, Sagot, and Dupoux (2020), this density estimation method was shown to correlate with true frequency of speech segments better than classical clustering methods based on k-means or hierarchical k-means.

The third step, as shown in Fig. 1, is to use these frequencies to compute the probability of each speech interval to be a word. To do that, DP-Parse relies on a Dirichlet Process formulation inspired by Goldwater et al. (2009). The details of the probability formulation are given in Algayres et al. (2022). The fourth step is to use these probabilities to estimate the ideal parse of each utterance. The ideal parse was defined as the one for which the multiplicative product of the probabilities of the constituent segments was maximized. Because the sentences were long, there were many possible parses, each made up of a sequence of

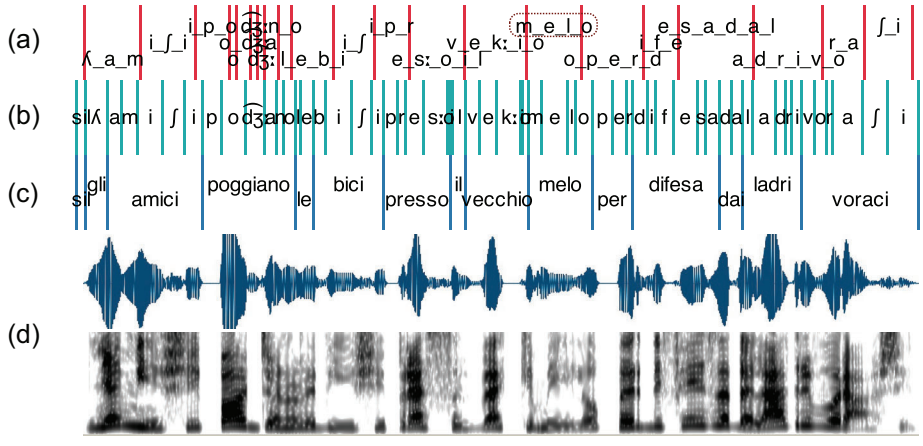

Fig. 2. One sentence from the Pelucchi et al. (2009a) corpus, together with one model output and gold-standard annotation. (a) is the DP-Parse output intervals in one run; (b) is the gold-standard phone parse; (c) is the gold-standard lexical parse; (d) is the waveform (amplitude over time) and spectrogram (energy in frequency bands over time). Successful segmentation of the word *melo* is highlighted with a dashed oval.

prospective words. Computing the full space of probabilities for every possible parse is not feasible, so an estimate of the best parse is obtained using an efficient implementation of the beam search algorithm based on dynamic programming. This is done by creating a lattice of the different parses for each sentence, and computing the most likely segmentations.

As a fifth and final step, the intervals making up this prospective "best" parse are then used to update the instance-based lexicon. The log-probability of the full corpus is obtained by summing the log-probabilities of all segmented sentences. The pipeline cycles in this fashion until the total corpus log-probability stops increasing.

The output of the model is, for each corpus, a sequence of time intervals, where each input sentence is exhaustively segmented. That is, apart from the sentence boundaries, the end of each interval is the start of the next one. The model does not output word estimates per se; it outputs intervals. Thus, the model did not produce a lexicon made up of discrete types, which may then be said to be present or not present in a given speech sequence. This is not to say that the model is incompatible with forming a lexicon; rather, doing so would be an add-on process based on the current outputs.

For the purpose of evaluating the intervals that the model produced, we lined up the model's intervals with our own hand-coded gold-standard phone alignments. An interval was credited with a phone if it overlapped with that phone's gold-standard interval for at least 30 ms. (This means that a phone token could be present in more than one model output interval.) These phone sequences were then evaluated to check for matches to the target lexicon of {bici, casa, fuga, melo}. An example of this alignment is shown in Fig. 2. Intervals output by one run of the DP-Parse model are given in the top row (see caption for details).

Thus, here we started from the audio sentences of each Pelucchi corpus, trained speech embeddings for these materials, and segmented the sentences using DP-Parse. The model

was run 100 times over each of the four corpora of sentences, two from each of the published papers: the experiment 3 stimuli of Pelucchi et al. (2009b), that is, Languages 3a and 3b, and Languages A and B from Pelucchi et al. (2009a).

We wanted to test whether a system that is not explicitly given phones or a lexicon to work with could successfully yield segmentation results similar to those implied by the infants' results. Although there are elements of the model's functioning that are unlikely to match infants' true mental processes (in particular, the model's iterative cycling through candidate parses), the capacity of a model to produce outputs that align with infant behavior (or the cognitive outcomes assumed to underlie that behavior) can be viewed as a minimal threshold a model should achieve. That is, our system succeeded when it was more likely to extract high-probability bisyllables than low-probability ones. As we will see, this was usually true. At the same time, it was not *always* true, and overall, recall rates were not very high. Our primary questions in analyzing the results were:

1. Considering the output intervals that matched a target word, were these HTP words in the corpus more often than LTP words?
2. Within the HTP and LTP conditions, were certain instances more likely to be found than others?
3. Did "near misses," where a model interval contained the target but also an additional phone or two, favor HTP words over LTP words?
4. Did DP-Parse find the "decoy" syllables that would help explain the lower discovery of LTP words relative to HTP words?

2. Analyses

First, we want to know if the model pulled out exact word matches more often for HTP than LTP words. In each corpus, there was a maximum of six matches to each of the four words. Infants in each of the modeled experiments listened longer, on average, to the HTP words (when those words were presented in isolated-word lists). The papers do not report separate statistics for each word, apart from saying that the results were similar across passages within conditions. Was this true for the model outputs?

Yes: it was (Fig. 3). At the level of the corpus (left panel), the model was more likely to pull out the HTP words than the LTP words, in four of four instances. At the level of the word (right panel), there was some variability: the HTP words tended to be extracted more often than the LTP words, but in the experiments of Pelucchi et al. (2009b), the HTP words did not dominate every time, within experiment. If we consider the words across test languages, the HTP instances of each word were found more often than their LTP instances in every case but one, namely, *melo* in the experiments of Pelucchi et al. (2009a).

Second, which instances were easy or hard? Or were the different instances of a word within a corpus similar in their likelihood of being found by DP-Parse? We can get a sense of that from Fig. 4.

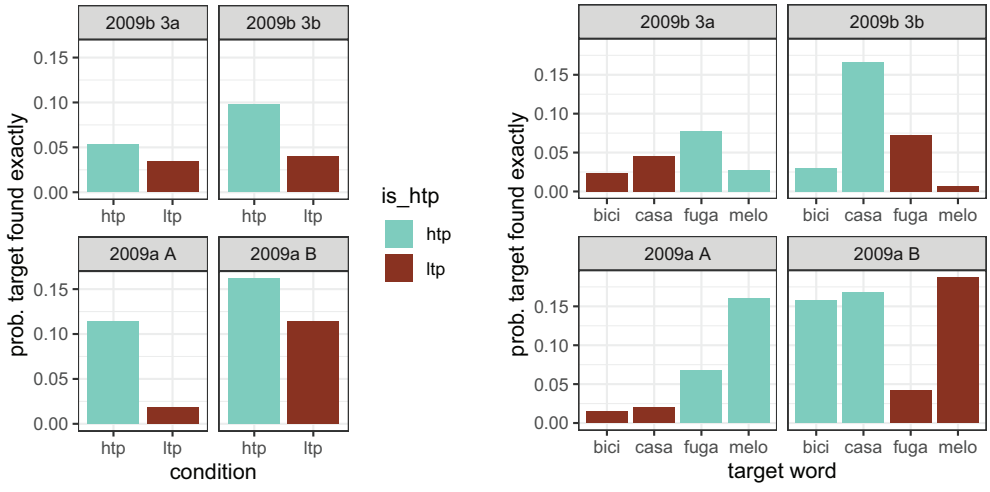

Fig. 3. The proportion of times a word was extracted, exactly, out of the times it was present in the corpus. The left panel averages over the target words, presenting averages by condition; the right panel shows results for each target word. The 2009b study's ltp words had low forward transitional probabilities; the 2009a study's ltp words had low backward transitional probabilities.

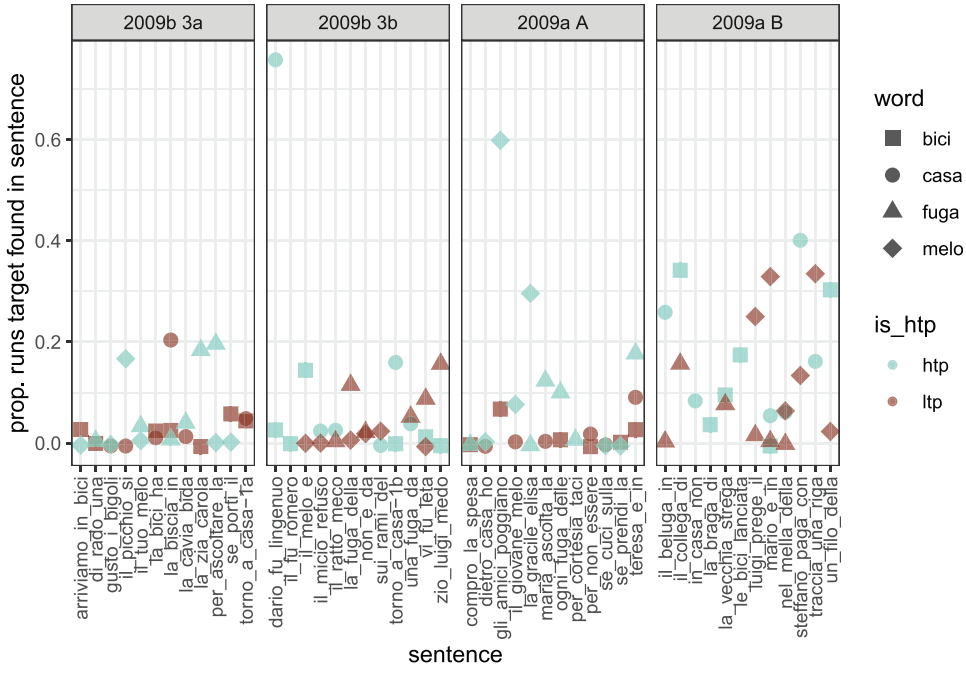

Fig. 4. Exact discovery probability by sentence, each word, faceted by corpus. Sentences are indicated according to their first three words (see Supporting Materials for the full transcripts).

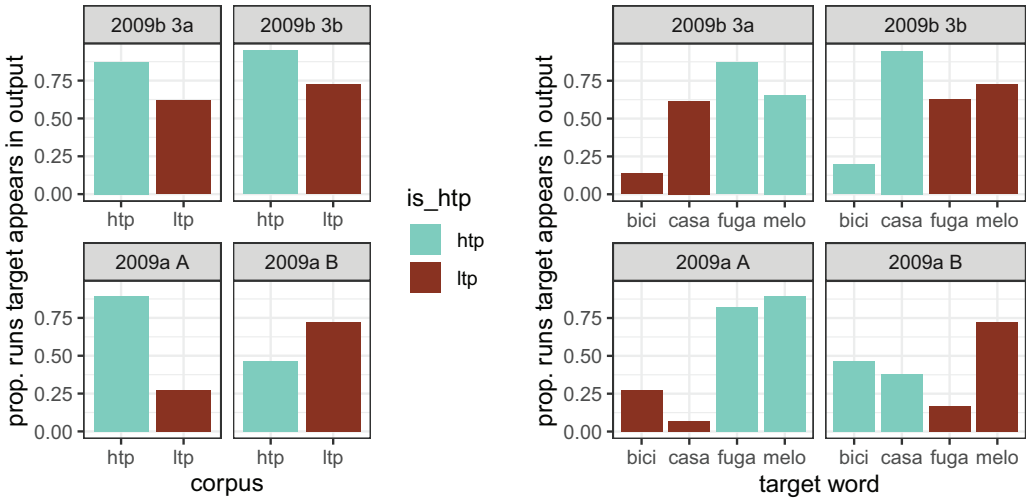

Fig. 5. The proportion of times a word was extracted, out of the times it was present in the corpus, allowing for extra phones on either side of the target. The left panel averages over words; the right panel shows results for each word type.

In the plot, the sentences are arrayed in columns, with each facet corresponding to a corpus. Each column has one, two, or three symbols on it, reflecting the fact that sentences in the stimulus passages varied in how many target words were present in them. What we expect to see is that for each corpus, HTP words would be found more often (higher on the y-axis) than LTP words. This was transparently true for Pelucchi et al. (2009a) language A, and less obvious for the others, though the results presented earlier did show the expected overall advantage for HTP words. Perhaps the most surprising result is that most sentences yielded a proportion of zero. That is, for most sentences, DP-Parse did not ever find the word embedded in it. Whether this is true of infants is unknown. We have not yet evaluated rigorously what it is about a token that makes it easy or difficult to extract exactly, but informally, listening to the sentences where words were easier to segment did not suggest to us any obvious characteristics of the more extractable words.

Third, are the results similar for "near misses," where DP-Parse extracted a target, but with extra material at the start or end? We can begin by looking at the probabilities of extracting a word at all, in any length interval, as opposed to splitting a word between intervals. Plots of these outcomes are given in Fig. 5.

These results are similar to the exact-matches case, except that the HTP advantage has gone away, in the case of the 2009a Language B corpus. This seems to be because *melo* in that corpus was easy to find. We will see a partial explanation for this later.

What if we are more stringent about partial matches than simply requiring that an entire word appear somewhere in an interval? To examine this, we can analyze as hits any DP-Parse output interval that contained a target word plus one phone (such as *a.bici* or *bici.l*, for

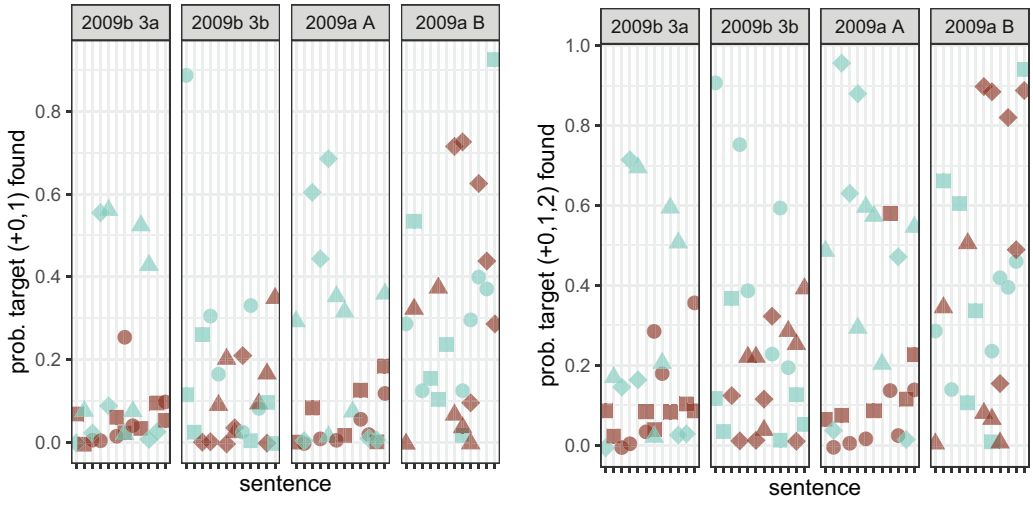

Fig. 6. The proportion of times a word was extracted, out of the times it was present in the corpus, allowing for one or two (total) extra phones on either side of the target. First plot (left): zero or one extra phones; second plot (right): zero, one, or two extra phones. Legend is as in the previous figure.

instance), or one or two phones, either both at the start or end, or one in each spot. If we consider these as hits, the distribution over sentences is given in Fig. 6.

For the most part, this view of the data appears to enhance the advantage of the HTP words, or at least maintain it. However, in 2009a Language B, *melo* did surprisingly well as an LTP word.

Fourth, did DP-Parse detect syllables? Just as we evaluated the learner's intervals that contained exactly a target word, we can do the same for intervals that contained exactly one of the eight consonant-vowel syllables that composed the target words. Here, we look at the number of times an output interval was identical to a given target syllable, divided by the number of times that syllable appeared in the corpus. Note that this denominator is much larger in the LTP case, for one of the two LTP words' syllables. In the Pelucchi et al. (2009b) corpora, the LTP words' first syllables occurred an additional 12 times in other words; in the Pelucchi et al. (2009a) corpora, the LTP words' second syllables occurred an additional 12 times. The plot accounts for this, though, in presenting recall probabilities, that is, counts of hits divided by the number of possible hits. See Fig. 7.

In general, the syllables from LTP words were discovered much more often as entire intervals than the syllables from HTP words were. As predicted by the design of the infant experiments, the "decoy" syllables were frequently treated as their own units. In the case of the 2009b studies (first and second facets), the frequent decoys were the stressed syllables (*bi, ca, fu, me*), represented on the plots by the larger plotting symbols. They were pulled out often. In the 2009a studies, the frequent decoys were (*ci, sa, ga, lo*), the unstressed second syllables of the target words. These were found sometimes (Lang A: *ci* especially, sometimes *sa*; Lang B: *lo*). What is interesting in Language B is how often *bi* and *ca* were discovered as units,

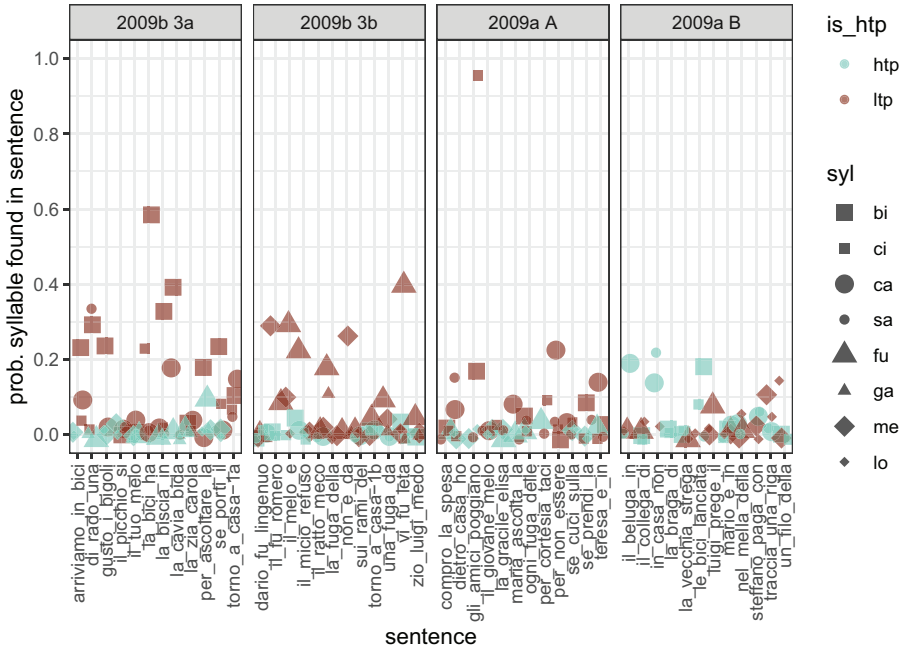

Fig. 7. The proportion of times a syllable was extracted, exactly, out of the times it was present in the corpus.

even though in every instance those syllables were the initial parts of full HTP target words. In addition, the decoy units in Language B, *ga* and *lo*, were found relatively infrequently. This may help explain why *melo* was extracted as a word so often from the Language B corpus, contrary to expectations.

3. "Weak Supervision" results

Thus far, the results presented have all been from the maximally unsupervised version of the model. However, it is possible to provide the model with more training on the speech sounds in the corpus, as described in the Introduction. Although this training has some supervision, and, therefore, could not strictly align with learning done during the experiment, it does provide a way to simulate the hypothesis that infants are better at adapting to a new talker (and here, a new accent and new language) more readily than the self-supervised model is. The weakly supervised speech model of DP-Parse was provided with additional contrastive training. In this training, speech chunks of random duration were pulled out from the training corpus of a given simulation and paired up with other such speech chunks that contained the same gold-standard phones. Training served to enhance the representational similarity of all pairs that shared speech-sound labels, and decrease the representational similarity of all mismatching pairs (the large majority). Note that this training boost did not have access to exact phone boundaries, only the string of phone labels in each random chunk.

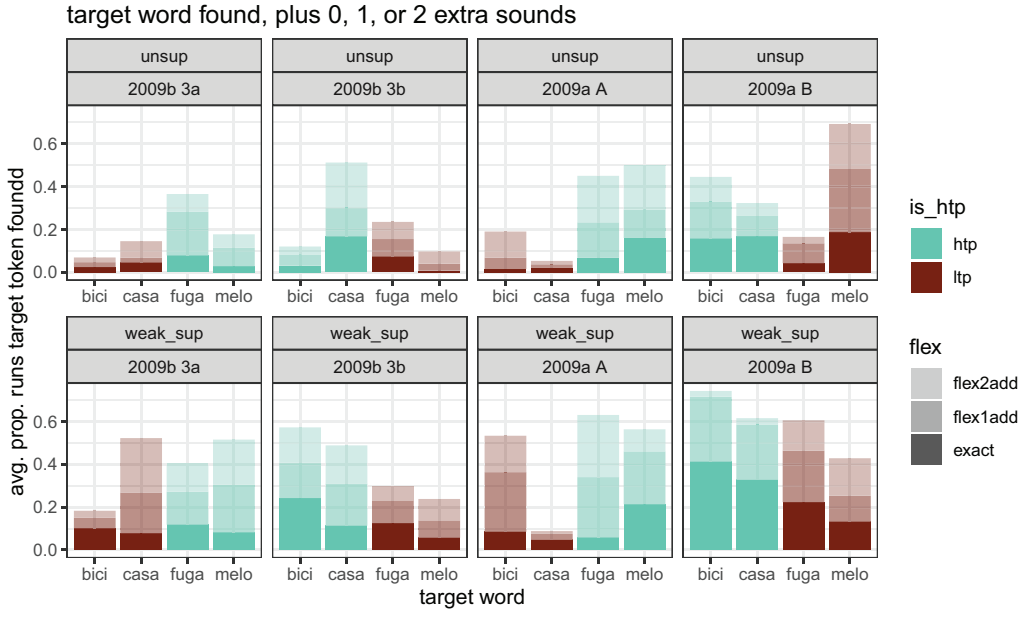

Fig. 8. The proportion of times each target was extracted, possibly with intervals that include a small number of extra sounds, out of the times it was present in one run of the corpus, for self-supervised training (top row) and weakly supervised training (bottom row). "Exact" refers to intervals containing only the canonical sounds of the targets; "flex1add" refers to additional probability gained by allowing as hits intervals including one extra sound either at the start or the end; "flex2add" refers to additional probability gained by allowing a total of two extra sounds, such as one at the start and one at the end.

This "weak supervision" had the effect of making target words more likely to be found. For two corpora (2009b, set 3b; 2009a, set B), the difference between HTP and LTP targets was larger; for two corpora (2009b, set 3a; 2009a, set A), the difference was smaller. In no case did the effect invert when the model was given more training.

These results for exact matches are shown together with analogous results for more flexible matching (the word plus a phone or two) in Fig. 8. Outcomes for models with self-supervised training are shown in the top row, with exact matches presented as the darkest bars and near-misses as progressively lighter bars. Outcomes for models with weak supervision are shown in the bottom row.

How should we interpret the proportion of targets found (displayed on the y-axis)? The self-supervised case shows average likelihoods of 0.11; the Pelucchi et al. (2009b) experiment 3a, in which infants did indeed succeed in recognizing HTP words, reveals a recall proportion of only 0.053. Each target word appeared in each corpus six times, and each corpus was played through three times. Thus, an average recall rate of 0.053 implies only an average of 18 * 0.053 or 0.95 tokens being retrieved and, therefore, forming the basis for the infants' response. Is this enough to drive a response? Intuitively, the number feels low. It is possible that the (presumed) feeling of familiarity that drives infants' looking-times would also be supported

by near-misses of the word, that is, the segmentations that we noted in Figs. 4 and 6. In studies involving familiarization and test procedures, infants do not respond to phonological near-misses as if they were correct canonical tokens (e.g., Jusczyk & Aslin, 1995; Mattys & Jusczyk, 2001), but if provided a mixture of canonical tokens and near-misses, the near-misses might boost familiarity.

4. Conclusions

We find that the DP-Parse model approximated the speech processing implied by the results of the infant experiments conducted by Pelucchi and colleagues. Although the fact that the methods employed in those experiments do not yield item-level information prevents us from making a more detailed comparison of what infant and machine extracted from the passages, we believe that comparisons of this sort can be useful, in principle, in the evaluation of computational models of infant language processing.

Why was the DP-Parse model more likely to extract bisyllabic target words when they had an HTP in the corpus, given that DP-Parse does not compute transitional probabilities? DP-Parse, like the DP-Unigram model, favors intervals that occur frequently. If a syllable like *bi* occurs many times in a language sample, it will be favored as a unit, including within the word *bici*. Effectively, then, we might say that it is not so much that the HTP targets are preferred, but that the LTP targets lose their nominally high frequency of occurrence via the theft of their more frequent component syllables. A word like *bici* might occur six times in a series of sentences, but if it is usually experienced as *bi* + {something else}, the six instances do not register as such. This happens in the parsing stage represented by the lattice at the bottom of Fig. 1. If *bi* has been discovered as a frequent unit, {bi}{ci…} becomes more probable; if *bici* is common relative to *bi* or *ci*, then {bici} becomes more probable. This basic mechanism is conceptually similar to the mechanism of some other computational models that have similarly argued that an apparent sensitivity to conditional probability variation could emerge without actual computation of conditional probability per se (Cabiddu et al., 2023; Perruchet & Vinter, 1998; see also Wang, Hutton, & Zevin, 2019). Denying that explicit computation of syllabic transitional probability statistics is necessary for statistical learning leaves open several fundamental questions about the nature of the learning. We consider it likely that parsed tokens become chunks that are mentally represented as such, and that may become *protolexical* units available for entry into syntactic and semantic linguistic networks if they continue to be supported in further language experience (e.g., Swingley, 2007). But how and when this happens is a matter of debate, and may involve multiple neurally distinct processes (e.g., Henin et al., 2021; Sučević & Schapiro, 2023; Wang et al., 2019). The modeling presented here is neutral in this regard.

Any computational model meant to mimic the developmental progress of the infant's mental representations for language should, ideally, acquire the same knowledge from experimental stimuli that actual infants apparently do, at least as inferred from the logic of preferential listening experiments. Testing models on experimental stimuli is different from the more typical evaluation in which a model's outputs are scored according to gold standards given by

assumptions about the language—its set of consonants and vowels, the canonical representations of its words, and so on. One risk of the latter sort of evaluation is that it might project onto infants representations they do not actually possess at the developmental stage being modeled. Evaluations based on outputs from exposure to experimental materials bring the test closer to what can be claimed empirically about infants' knowledge state.

The set of infant word segmentation experiments is small (on the order of several dozen), and the set of experiments for which auditory materials and exact transcripts are available is smaller still. The tests presented here required, as a preliminary step, the segmentation of the stimulus corpora into a gold-standard set of phonological and lexical transcriptions. This is a nontrivial process. Still, we argue that in the long run, evaluating computational models against a battery of speech materials from infant studies would be a productive testing strategy, and recommend that future studies of infants' treatment of speech materials provide those materials freely for use in the benchmarking of computational models, as we have begun to do here. Ultimately, as models become better able to mimic infants' learning in these experiments, comparisons between models will drive the design of more refined experiments, ideally with the ability to place firmer quantitative constraints on estimates of infants' abilities. In this way, it should be possible for our field to move toward models that can make more fine-grained quantitative predictions about development.

Author contributions

The authors made the following contributions. Daniel Swingley: Conceptualization, project administration, software, supervision, validation, writing—original draft preparation, writing—review and editing; Robin Algayres: Conceptualization, project administration, software, writing—review and editing.

Acknowledgments

Thanks are due to Dr. Jessica Hay for her help in providing the stimulus files and transcripts of them.

Notes

1 The 40 ms grain size is intended to capture the rapid change typical of continuous speech without blowing up the amount of computation to run the model. Although we have not explored varying this parameter, we suspect that a longer grain size, say 80 or 100 ms, would too frequently grossly misalign with syllable or word boundaries, blurring similarities we need to represent.
2 In this paper, we follow the speech technology usage of "segment" referring to any portion of the speech signal, rather than the linguistics usage of "segment" referring to a consonant or vowel.

References

Algayres, R., Nabli, A., Sagot, B., & Dupoux, E. (2023). *Speech sequence embeddings using nearest neighbors contrastive learning*. Retrieved from https://arxiv.org/abs/2204.05148

Algayres, R., Ricoul, T., Karadayi, J., Laurençon, H., Zaiem, S., Mohamed, A., Sagot, B., & Dupoux, E. (2022). DP-Parse: Finding word boundaries from raw speech with an instance lexicon. *Transactions of the Association for Computational Linguistics*, *10*, 1051–1065.

Algayres, R., Zaiem, M. S., Sagot, B., & Dupoux, E. (2020). *Evaluating the reliability of acoustic speech embeddings*. Retrieved from https://arxiv.org/abs/2007.13542

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Beckman, M. E., & Edwards, J. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development*, *71*, 240–249.

Beech, C., & Swingley, D. (2023). Consequences of phonological variation for algorithmic word segmentation. *Cognition*, *235*, 105401.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724.

Bergelson, E., & Swingley, D. (2012). At 6 to 9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the USA*, *109*, 3253–3258.

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, *52*, 264–278.

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2023). CLASSIC utterance boundary: A chunking-based model of early naturalistic word segmentation. *Language Learning*, *73*(3), 942–975.

Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T. A., Seyssel, M. de, Rozé, P., Rivière, M., Kharitonov, E., & Dupoux, E. (2021). The Zero Resource Speech Challenge 2021: Spoken language modelling. *CoRR, abs/2104.14700*. Retrieved from https://arxiv.org/abs/2104.14700

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 184–193).

Feldman, N. H., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do infants really learn phonetic categories? *Open Mind*, *5*, 1–19.

Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics*, *53*, 287–295.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Hallap, M., Dupoux, E., & Dunbar, E. (2023). *Evaluating context-invariance in unsupervised speech representations*. Retrieved from https://arxiv.org/abs/2210.15775

Hallé, P. A., & de Boysson-Bardies, B. (1994). Emergence of an early receptive lexicon: Infants' recognition of words. *Infant Behavior and Development*, *17*, 119–129.

Harwath, D., Torralba, A., & Glass, J. (2016). Unsupervised learning of spoken language with visual context. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 1–9). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2016/file/82b8a3434904411a9fdc43ca87cee70c-Paper.pdf

Henin, S., Turk-Browne, N. B., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., Devinsky, O., & Melloni, L. (2021). Learning hierarchical sequence representations across human cortex and hippocampus. *Science Advances*, *7*(8), eabc4530.

Jusczyk, P. W. (1993). From general to language-specific capacities: The Wrapsa model of how speech perception develops. *Journal of Phonetics*, *21*, 3–28.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.

Kamper, H., Jansen, A., & Goldwater, S. (2016). Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *24*(4), 669–679.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.

Mattys, S. L., & Jusczyk, P. W. (2001). Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 644–655.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494.

Matusevych, Y., Schatz, T., Kamper, H., Feldman, N. H., & Goldwater, S. (2023). Infant phonetic learning as perceptual space learning: A crosslinguistic evaluation of computational models. *Cognitive Science*, *47*(7), e13314.

Mnih, A., & Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. *arXiv Preprint arXiv:1206.6426*. https://doi.org/10.48550/arXiv.1206.6426

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, *33*(3), 1065–1076.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244–247.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*, 674–685.

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, *122*(4), 792.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*, 113–146.

Settle, S., & Livescu, K. (2016). Discriminative Acoustic Word Embeddings: Current Neural Network-Based Approaches. *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning. *eLife*, *12*, e77185.

Swingley, D. (1999). Conditional probability and word discovery: A corpus analysis of speech to infants. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 724–729). Mahwah, NJ: LEA.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86–132.

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, *43*, 454–464.

Swingley, D. (2021). The identifiability of consonants and syllable boundaries in infant-directed English. *Paper Presented at the 34th Annual CUNY Conference on Human Sentence Processing*. Philadelphia.

Vihman, M. M., & Croft, W. (2007). Phonological development: Toward a 'radical' templatic phonology. *Linguistics*, *45*, 683–725.

Wang, F. H., Hutton, E. A., & Zevin, J. D. (2019). Statistical learning of unfamiliar sounds as trajectories through a perceptual similarity space. *Cognitive Science*, *43*(8), e12740.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental model of speech processing. *Language Learning and Development*, *1*, 197–234.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Yates, T. S., Skalaban, L. J., Ellis, C. T., Bracher, A. J., Baldassano, C., & Turk-Browne, N. B. (2022). Neural event segmentation of continuous experience in human infants. *Proceedings of the National Academy of Sciences*, *119*(43), e2200257119.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Materials