J. Child Lang. **37** (2010), 513–543. © Cambridge University Press 2010
doi:10.1017/S0305000910000085

Segmenting words from natural speech: subsegmental variation in segmental cues*

C. ANTON RYTTING

University of Maryland Center for Advanced Study of Language (CASL) and Department of Linguistics, the Ohio State University

CHRIS BREW AND ERIC FOSLER-LUSSIER

Department of Computer Science and Engineering and Department of Linguistics, the Ohio State University

(*Received* 20 *December* 2008 – *Revised* 29 *October* 2009 – *Accepted* 30 *January* 2010 – *First published online* 22 *March* 2010)

ABSTRACT

Most computational models of word segmentation are trained and tested on transcripts of speech, rather than the speech itself, and assume that speech is converted into a sequence of symbols prior to word segmentation. We present a way of representing speech corpora that avoids this assumption, and preserves acoustic variation present in speech. We use this new representation to re-evaluate a key computational model of word segmentation. One finding is that high levels of phonetic variability degrade the model's performance. While robustness to phonetic variability may be intrinsically valuable, this finding needs to be complemented by parallel studies of the actual abilities of children to segment phonetically variable speech.

INTRODUCTION

One of the fundamental questions in child language acquisition is how children learn to segment running speech into words. Children's performance at segmenting words is a good predictor of later performance at other stages of language learning, including not only vocabulary building but also

[*] Portions of this research were conducted with the monetary support of a National Science Foundation Graduate Research Fellowship awarded to the primary author while he was at the Ohio State University, as well as from NSF-ITR grant #0427413, granted to Chin-Hui Lee, Mark Clements, Keith Johnson, Lawrence Rabiner and Eric Fosler-Lussier for the multi-university Automatic Speech Attribute Transcription (ASAT) project. Preliminary versions of parts of this work, in particular Simulation 1, appear in the primary author's (unpublished) dissertation. Address for correspondence: C. A. Rytting, 7005 52nd Avenue, College Park, MD 20742, USA. tel: +1 (301) 226-8883. e-mail: crytting@casl.umd.edu

expressive language and general language ability (cf. Newman, Bernstein Ratner, Jusczyk, Jusczyk & Dow, 2006).

Most computational models of the word segmentation task treat it as a series of subtasks along these lines:

1. Break the stream of speech into a sequence of acoustic segments.
2. Convert these acoustic segments into abstract symbols or feature bundles.
3. Find coherent subsequences within this sequence of symbols. Treat these subsequences as 'candidate words'.
4. Add these candidate words into a mental lexicon, combining identical subsequences into the same lexical entry, as instances of the same word.

Most models take steps 1–2 as a given (or as a necessary idealization), and focus on step 3, using transcriptions produced by adults (or dictionaries) as the source of the input. An alternative way of conceptualizing the word learning task involves a slightly different series of subtasks:

1. Break the stream of speech into a sequence of acoustic segments.
2. Represent these segments in a way that preserves the acoustic variation in each segment, such as probability vectors over possible symbols.
3. Find coherent subsequences within this sequence of probability vectors (PVs). Treat these subsequences as 'candidate words'.
4. Add these candidate words into a mental lexicon, clustering very similar subsequences of PVs into the same lexical entry as likely instances of the same word.

We have developed and implemented a framework for providing alternative input to any computational model of word segmentation that can accommodate probabilistic input. This alternative input follows steps 1–2 as we propose them, using techniques from automatic speech recognition as a proxy for human auditory processing. We argue that this input is more realistic than the types of transcripts usually used to train and test computational models of word segmentation. We also demonstrate the use of this new input on one model of word segmentation (Christiansen, Allen & Seidenberg, 1998).

The paper is organized as follows. First, we review previous work modeling subsegmental variation and propose a method of preserving the phonetic variation found in audio input that is compatible with well-established evaluation metrics, and hence readily comparable across models. We then describe how the input data derived from this method is applied to the Christiansen *et al.* (1998) model of word segmentation. Two simulations follow: the first simulation tests the generality of the claims in Christiansen *et al.* (1998), by comparing the model's performance using symbolic, citation-form input against that using probabilistic, speech-derived input. To simplify the comparison, a version of the model without suprasegmental

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

cues was used. The results of this simulation show that the chosen word segmentation model performs well given input with little phonetic variation, but not with the high levels of acoustic variation found in many recorded utterances in the Brent corpus (Brent & Siskind, 2001).

The second simulation tests the generality of the claims about how best to combine multiple cues. It compares two variants of the model that combine segmental information with a novel measure of speech clarity called SEGMENTAL SALIENCE. This cue is used as a rough approximation to suprasegmental cues such as lexical stress, which are known to be of importance in word segmentation for English (Johnson & Jusczyk, 2001). The second simulation confirms that multiple cues can, under certain conditions, be combined to improve word segmentation performance. It also indicates that phonetic variation cannot be ignored when designing studies of the statistical learning of language. In addition, it suggests that more refined evaluation measures will be needed if we are to understand the advantages and disadvantages of particular statistical learning techniques.

Finally, we discuss the implications for other word segmentation models that rely heavily on segmental information, as well as opportunities for investigating word segmentation performance in non-ideal circumstances. Whether children are able to extract word boundaries from the more highly variable utterances in the Brent corpus is not clear; this warrants further investigation. As computational models of word segmentation investigate performance on concrete examples of running speech, more coordination with experimental measures of infant performance will be necessary to interpret the results.

RELATED WORK: MODELING SUBSEGMENTAL VARIATION

Computational models of infant language acquisition only rarely represent the subsegmental variability present in speech. Those models that have examined the effect of phonetic variation have used two approaches: first, the use of automatic speech recognition (ASR) technology, and second, stochastic resetting of phonological features.

Using automatic speech recognition

Carl de Marcken (1996) was perhaps the first to use ASR technology in modeling the word segmentation task. In order to avoid including information unavailable to the infant, de Marcken's model used the output of an automatic phone recognition (APR) system, explicitly excluding all higher-level linguistic information that would bias the system towards phone sequences more frequently found in canonical pronunciations of words. However, de Marcken's experiments used Viterbi's algorithm to

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

reduce speech to a single most likely sequence of phones. Thus, de Marcken's approach does no more to model uncertainty or ambiguity at the segmental level than phone-level human-transcribed corpora, such as the Buckeye corpus (Pitt, Johnson, Hume, Kiesling & Raymond, 2005) used by Fleck (2008). Both over-commit to the phonemic identity of the segment before segmentation begins. The difference is that de Marcken replaces human phone recognition error with APR error.

Roy & Pentland (2002) do handle uncertainty: their CELL model is similar in many respects to the approach proposed and tested in this paper. However, since the CELL model focuses primarily on the task of word–meaning mapping within a multimodal domain, its performance was never compared to any other word segmentation model, nor was it tested on comparable corpora. Neither de Marcken (1996) nor Roy & Pentland (2002) provide evaluation metrics that can be directly compared with other models.

Simulating phonetic variation

Two connectionist approaches, Cairns, Shillcock, Chater & Levy (1997) and Christiansen & Allen (1997), simulate subsegmental variation in the input by stochastically inserting non-canonical combinations of phonetic features. The latter study is described in more detail below.

The Christiansen & Allen (1997) study (henceforth CA97) used the Carterette & Jones (1974) corpus as input, encoding each phoneme as a vector of (binary) phonological features. During testing, certain of these input features were randomly flipped from 0 to 1 or from 1 to 0, in order to test the effect of subsegmental variation on the connectionist model's performance. Those features which distinguished a particular phoneme from another American English phoneme were considered CORE features and left unchanged. Of the rest, certain features (on average about two per phoneme) were dubbed PERIPHERAL features and were randomly and independently toggled at various probabilities (four conditions: 0, 0·01, 0·05 and 0·1) in order to simulate subsegmental variation in the speech signal. It was found that this type of variation did not significantly alter the performance of the neural network, either when starting with citation forms or the (human) phone-level transcriptions of the corpus.

CA97's implementation of subsegmental variation, while ingenious, is somewhat ad hoc. The distinction between core and peripheral features amounts to an implicit theory about segment confusability. As a theory, this leaves something to be desired, underestimating the probability of confusions that seem quite plausible for infants without top-down information. For example, since English obstruents are arranged in voiced–voiceless minimal pairs (e.g. /f/∼/v/, /t/∼/d/, /s/∼/z/), the feature [voice] would be a core feature for all obstruents under CA97's assumptions. This would mean that

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

it is impossible to misclassify any obstruent along the line of voicing (e.g. a /t/ for a /d/). In reality, voice-onset time is continuous, and the slope of human categorical perception, while steep, is known not to be absolutely vertical (cf. McMurray & Aslin, 2005), so two slightly differently tuned listeners certainly could perceive the value of the [voice] feature differently for a particular phone, particularly one near the VOT (voice-onset time) boundary for voicing. By not allowing for this possibility, CA97 effectively models the [voicing] feature as a perfect step function for obstruents. While this is a retreat from one kind of idealization, it also introduces unlooked for and cognitively unmotivated assumptions that should perhaps be eschewed.

Second, the model controls variation in each peripheral feature with a single number that is constant across time and surrounding context. This has the effect of spreading the subsegmental variation (or phonetic ambiguity) in the model evenly throughout each word and each utterance. Again, this is cognitively implausible: surrounding context, including position in the syllable (e.g. Redford & Diehl, 1999) and surrounding phones (e.g. Krull, 1990), plays a large role not only in allophonic variation but also in perception of phonemic distinctions and specific patterns of confusability.

Finally, infants experience subsegmental variation and phonemic ambiguity (and hence the possibility of phonemic confusion) throughout the course of acquisition. CA97 treats variation only at test time. This corresponds to a setting where the learner has access to very clear input (perhaps from a very cooperative caregiver) during training but must handle variability (maybe from other speakers) at later points. Unfortunately, speech science does not support the claim that even the clearest speech is free of variation. For prediction tasks like the one used in CA97, the use of subsegmental variation during training should apply to the target layer as well as to the input layer, since the training cannot presuppose any more certainty than the original input.

Now that high-quality audio data are available, a more direct and straightforward way of modeling phonetic variation is possible. We do not have to build a theory of phonetic variation, but can simply use the subsegmental variation already present in speech. We do this while retaining a model and metrics that are comparable to previous word segmentation models.

MODELING SUBSEGMENTAL VARIATION WITH PROBABILISTIC INPUT

Phone probability vectors

Most models of the word segmentation task make the following assumptions, or closely related variants:

1. The input provided to the model is represented as a sequence of symbols – that is, a string drawn from a finite alphabet or phonemic

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

inventory, augmented by a symbol for utterance boundaries. In the case of distributed (connectionist) representations, each combination of features maps to a symbol.

2. Only one label (or combination of features) is associated with each position in the input string.
3. The number of positions (or segments) in the input string is fixed: each utterance has a certain, known number of segments.
4. The word segmentation task is framed either as (a) the task of identifying which pairs of adjacent segments should have word boundaries placed between them, or (b) the task of identifying which subsequences of contiguous segments (substrings) should be grouped together as words.[1]
5. Evaluation consists of measuring the accuracy of (a) boundary placement, (b) substring groupings (or word tokens), and/or (c) the building of a lexicon of distinct substrings (or word types).[2]

We propose a different set of assumptions, which we argue are more realistic representations of the input infants receive. Specifically, we replace assumptions 1 and 2 with the following:

1′. The input provided to the model is represented as a sequence of probability vectors over a finite set of symbols (or PHONESET), augmented by a symbol for utterance boundaries.

2′. Many labels can be associated probabilistically with each segment of the input string, as long as the total probability over all the labels sums to 1. The labels and probabilities associated with a segment of the input string constitute that segment's PHONE PROBABILITY VECTOR.

For the sake of maintaining the evaluation metrics commonly used in other models, and facilitating comparison with them, we keep assumptions 3 through 5 constant. Along with assumption 3, we further assume that each segment in an utterance is associated with a specific temporal region in that utterance's audio recording, and that there are no gaps or overlaps between adjacent segments. Finally, we frame the word segmentation task as a boundary-detection task (as in 4a, above) and evaluate the resulting

[1] If all boundary or grouping decisions are deterministic, then the results of these two approaches are interchangeable.

[2] CAS98 and other models propose and use additional evaluations as well; that being said, the evaluations listed here seem to be more general across the modeling community. Only a few models (e.g. Roy & Pentland, 2002) address the issue of which acoustic variants should be clustered into the same lexical entry – a problem which can arguably be grouped with a later word–meaning mapping task rather than word segmentation per se.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

segmentation by the three metrics listed in assumption 5, discussed in more detail below.

Automatic phone classification

Our input representation is sensitive to what Scharenborg, Norris, ten Bosch & McQueen (2005) refer to as 'probabilistic acoustic detail', but is more constrained than the phone lattice that they consider. In traditional ASR lattices, as in Scharenborg *et al.*'s description, competing phones may be of different lengths, and the task of the Viterbi algorithm is to choose the best sequence of phones to cover the acoustic input. The optimization process is a search that considers all reasonable possibilities for the number of phones, their identity and the positions of the phone boundaries. In our application, which needs to achieve comparability with CA97, this degree of generality cannot easily be accommodated. Instead we treat phone boundaries (hence also number of phones) as fixed. For each position in the segmented input, we use the relevant acoustic material to assign a posterior probability for each of the possible phone labels in the recognizer's repertoire. Where CA97 works with a sequence of phone labels that are assumed to be reliably known, we work with a sequence of probabilistic distributions over phone labels. If the speech were particularly clear and the phone classifier especially effective, these distributions might turn out to be sharply focused on the 'correct' single phones, but in the more prevalent cases where the signal is less helpful or the recognizer less effective the result will be a distribution that allocates significant posterior probability to several different phone labels.

One variant of automatic speech recognition (ASR) that is compatible with these assumptions is automatic phone classification (cf. Halberstadt & Glass, 1997). Automatic phone classification (APC), like the automatic phone recognition variant that de Marcken used, differs from standard ASR in the basic unit of recognition and the types of linguistic knowledge that it utilizes. In traditional ASR, the basic unit is the word, and the task is to identify the sequence of words most likely to have given rise to a particular audio signal. In addition to a phoneset and an ACOUSTIC MODEL describing the ranges of audio input associated with each phone in the phoneset, an ASR system has a PRONUNCIATION DICTIONARY (or vocabulary of words and their usual pronunciations) and a GRAMMAR to describe how those words are likely to be sequenced together.

In APR and APC, the basic unit is not the word but the phone. Hence, no pronunciation dictionaries or word-level grammars are used, because we cannot assume that babies have a vocabulary yet; the goal of word segmentation is to find the words in order to acquire a vocabulary. APC differs from APR in that the former assumes that the number of phones in an

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

utterance and their boundary points are known. Hence, for each phone position and its associated temporal slice of audio signal, discovering the phone's identity may be construed as a classification task. APR does not make this assumption, which poses problems for evaluation. While 'hard-decision' APR as used by de Marcken returns unambiguous phone boundaries that can then be mapped to word boundaries, in a 'soft-decision' APR system, since it has probabilistic rather than predetermined phone boundaries, the choice-points for word boundaries are probabilistic rather than a discrete set, which breaks assumption 3 and makes evaluation and comparison with transcription-based word segmentation models much less straightforward. For this reason, APC is adopted for producing the input representations for this study.[3]

Obtaining the phone probability vectors

The conversion of the raw audio input into the sequence of phone probability vectors needed for the connectionist model's input is conducted in two stages: find the phone boundaries and then generate the phone probability vectors within each boundary. Both stages are implemented by using a previously developed ASR system based on the hidden Markov model toolkit HTK (Young *et al.*, 2002). The first stage divides each utterance (using the utterance boundaries marked in the corpus) into discrete, one-segment time intervals. In the second stage, each time interval between two phone boundaries is treated as a separate 'mini-utterance' for purposes of phone classification. The HTK system is constrained to treat each mini-utterance as a single segment. More details concerning the implementation of both steps are given in Rytting (2007).

This method for calculating the phone probability vectors for each segment does not utilize all of the typical contextual knowledge of a typical ASR system (including lexical pronunciations and/or phonotactic grammars) since each segment is considered in isolation; the only linguistic knowledge provided is the segment boundaries and the set of acoustic phonetic models. Because of the lack of contextual knowledge, the ASR acoustic models will be less accurate than those of a state-of-the-art system, but will also preserve subsegmental variation in the signal, which is crucial to model the sorts of uncertainty that an infant listener might experience. By utilizing

[3] This assumption excludes from consideration one type of phonetic variation: namely, insertion and deletion of segments relative to the canonical standard. In this regard, phone-based transcriptions, whether human based like the Buckeye corpus or automatic like de Marcken's APR-generated corpus, model a type of variation missed by this input representation. While such variation could be included by using a phone lattice representation of the input as in Scharenborg *et al.* (2005), this would significantly complicate evaluation of the input for many word segmentation models.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

ASR models which possibly overestimate, rather than underestimate, the amount of variation seen by an infant, we do provide a significant challenge to the model; however, if a model is successful with this type of input then clearly it can handle smaller amounts of variation.

OVERVIEW OF SIMULATIONS

An overview of the Christiansen, Allen and Seidenberg model

In order to investigate the effects of realistic, speech-derived subsegmental variation on word segmentation, we have compared the performance of one influential model of word segmentation using both symbolic and probabilistic input. We focus here specifically on the multiple-cue connectionist model described in Christiansen, Allen & Seidenberg (1998; henceforth CAS98), which we will refer to generically as the 'Christiansen model'. While a number of other models could in principle have been adapted to allow for probabilistic input (see, e.g., Batchelder (2002) for a review, and Fleck (2008), Frank, Goldwater, Mansinghka, Griffiths & Tenenbaum (2007) and Goldwater, Griffiths & Johnson (2009) for more recent models and empirical evaluations), the simple recurrent network at the basis of the Christiansen model is relatively easy to implement with widely available neural network toolkits and straightforwardly accepts probabilistic input.

CAS98 examines the interaction of multiple cues in finding 'hidden structure' in a sequence of observations. It hypothesizes that infants, while performing their primary task of learning the meaning of the language input they hear (or see), also engage in the IMMEDIATE TASK of learning to predict observable linguistic events, such as the identity of the next phone, the level of emphasis given to next phone, and whether or not the current phone precedes an utterance boundary. The finding of hidden structure such as word boundaries is a DERIVED TASK that emerges as infants attend to immediate tasks with directly observable feedback. This view of the infant's task is similar to that assumed by Aslin, Woodward, LaMendola & Bever (1996), where it is hypothesized that infants could find word endings by trying to predict utterance endings, and extrapolating from those phones (or features) that predict upcoming ends of utterances.

The innovation that the Christiansen model makes is how multiple cues are combined. Building on Aslin *et al.*'s immediate-task paradigm, CAS98 adds additional immediate tasks as CATALYST tasks, and trains an Elman network on these other tasks simultaneously. As multiple prediction tasks are learned simultaneously by the same network, the combined training will constrain the network to find a better joint solution for derived task of detecting hidden structure such as word boundaries. Specifically, CAS98 demonstrates that combining segmental and suprasegmental cues allows for greater performance at the word segmentation task than either cue alone.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

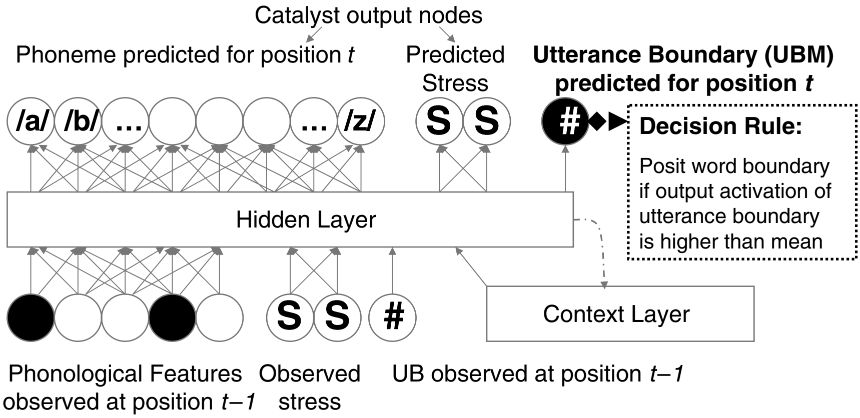
Catalyst output nodes
Phoneme predicted for position *t*
Predicted Stress
Utterance Boundary (UBM) predicted for position *t*
/a/
/b/
…
…
/z/
S
S
#
Decision Rule:
Posit word boundary if output activation of utterance boundary is higher than mean
Hidden Layer
S
S
#
Context Layer
Phonological Features observed at position *t–1*
Observed stress
UB observed at position *t–1*

Fig. 1. A schematic representation of the original CAS98 phon-ubm-stress network. Dark circles represent activated input units.

Figure 1 shows a schematic view of the Christiansen network. The catalyst output units for phonemic identity and stress constrain the model during training, but have no direct effect on the model's placement of word boundaries. Only the utterance boundary marker (shown by the # symbol, upper right) determines whether or not a word boundary is posited.

Goals of the simulations

In this paper we present two simulations illustrating the effect of probabilistic input on the Christiansen model. Simulation 1 examines the effects of probabilistic segmental input (i.e. phone probability vectors) on a version of the Christiansen network without lexical stress. Simulation 2 examines the impact of an additional cue (segmental salience). This additional cue is loosely analogous to suprasegmental cues such as lexical stress. Unlike Christensen's lexical stress cues, it is directly derivable from the phone probability vectors with no additional processing of the speech signal.

Implementation and evaluation of the Christiansen model

Model design, training and testing. In both simulations the Christiansen model was re-implemented using the Conx module of the Pyro toolkit (Blank, Kumar, Meeden & Yanco, 2003). Elman networks (a type of simple recurrent network or SRN) were used, following CAS98. Each network was trained on one pass through the training corpus, using the same settings as CAS98 (learning rate of 0·1, momentum of 0·95, and initial weight randomization ranging from −0·25 to 0·25), then tested with the network

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

weights 'frozen'. In order to account for the natural variability in the networks, nine separate runs of training and testing were performed for each variant of each network, each run differing only in the randomized starting weights.

Input representations. One point of difference between the original CAS98 study and the simulations described here is in the feature representation of the input to the connectionist model. Like Aslin *et al.* (1996) and Cairns *et al.* (1997), CAS98 represents each segment in the input corpus as a bundle of phonological features rather than as a discrete symbol. However, it is also possible to use symbolic input representations within a connectionist framework, by using a LOCALIST (or 'one-hot') representation, such that each symbol in the relevant phone set has one input unit uniquely associated with it. The original CAS98 study uses localist representations for their models' segmental output and target layers 'to facilitate performance assessments and analyses of segmentation errors' (p. 236).

We have conducted studies (not reported here) that examine the effect of input representation on the Christiansen model. For strict replication, we examined the original feature set in CAS98. This feature set contains some flaws (as pointed out in e.g. Fleck, 2008), so we also examined an arguably more realistic feature set, found in Christiansen, Conway & Curtin (2005). Finally, we examined a localist input representation matching CAS98's output and target layers. In general, the localist input representation performed as well as or better than either of the two distributed representations, so we report only the results of the localist representation here. The patterns observed from the distributed representations are not sufficiently dissimilar to affect the overall findings.

Evaluation procedures. Since the network is supposed to generalize from utterance boundaries to all word boundaries, the activation of the output unit corresponding to the utterance boundary marker (UBM) is used to determine the model's level of belief in a word boundary after a given segment. To calculate precision and recall (defined below), CAS98 posits a word boundary whenever the UBM output unit registers a greater-than-threshold activation. Following Aslin *et al.* (1996), the threshold used for determining a posited word boundary is the average activation for the UBM output unit over all positions. While this method of evaluation is not the only one provided by CAS98, it is the method most closely comparable to evaluations of other models in the literature, so we adopt it here in reporting the results of the Christensen model and its variants on new input.[4]

[4] Note that the boundary placement as calculated here considers both utterance-internal word boundaries and the final boundary at the end of the utterance, but not utterance-initial boundaries. A more conservative boundary measure, using only utterance-internal word boundaries, may be obtained by subtracting the number of utterances in the test corpus from the number of true positives (N_{tp}).

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

Results are reported in terms of PRECISION and RECALL (referred to as ACCURACY and COMPLETENESS in CAS98) for boundaries, word tokens and word types, where precision equals the proportion of true positives to the sum of true and false positives, and recall is the proportion of true positives to the sum of true positives and false negatives. Unlike CAS98, which only reports one run of the neural network, all simulations reported here take the mean values of true positives, false positives, and false negatives (i.e. $\langle \overline{N}_{tp}, \overline{N}_{fp} \rangle$ and $\langle \overline{N}_{tp}, \overline{N}_{fn} \rangle$) over nine separate runs with different (randomized) initial weights, as shown in Equations 1 and 2.

$$\textit{Mean Precision} = \frac{\overline{N}_{tp}}{\overline{N}_{tp} + \overline{N}_{fp}} \tag{1}$$

$$\textit{Mean Recall} = \frac{\overline{N}_{tp}}{\overline{N}_{tp} + \overline{N}_{fn}} \tag{2}$$

Following CAS98, significance in precision and recall between two conditions is measured comparing the mean number of true and false positives $\langle \overline{N}_{tp}, \overline{N}_{fp} \rangle$ for precision, and true positives with false negatives $\langle \overline{N}_{tp}, \overline{N}_{fn} \rangle$ for recall, for each of the two conditions in a 2×2 χ^2 test.

The number of boundaries correctly found is of less interest than the number of words (tokens and types) correctly segmented. In order for a word token to count as correctly segmented, three conditions must apply:

1. The word's beginning must be correctly identified.
2. The word's end must be correctly identified.
3. There must be no false-positive boundaries posited in between the beginning and end of the word.

The precision and recall over word types is calculated in the same manner as word tokens, except that each word type (or distinct string of canonical phones) is only counted once over the entire corpus. Type recall refers to the proportion of distinct words in the corpus found by the model (averaged over nine runs). Type precision refers to the proportion of distinct strings segmented and proposed by the model that correspond to actual words in the corpus, and corresponds to the 'lexicon precision' measure in Brent (1999). Since the Christiansen model does not compile a lexicon explicitly as part of its execution, the term 'type' precision is adopted here.

As with most computational models of word segmentation, no distinction is made in these metrics between different classes of words, such as function words vs. content words, nouns vs. verbs, or words that children show evidence of knowing vs. other words. Perfect recall means correctly segmenting ALL the words.[5]

[5] A more nuanced metric would undoubtedly be of interest for matching the performance of particular models with children's experimental performance at particular stages of word segmentation; however, no such metric has yet emerged as a community standard.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

SIMULATION 1

Although CA97 gives us some indication of the Christiansen model's performance in the face of certain types of variation in its input, the flaws discussed above in the section 'Simulating phonetic variation' limit the conclusions that can be drawn from it. Simulation 1 seeks to overcome these limitations by comparing the performance of the Christiansen model on citation-form input with its performance using phone probability vectors (derived from audio-recordings as described in the sections 'Modeling subsegmental variation with probabilistic input' above and 'Input data' below) as input for both training and testing. CAS98 used the Korman (1984) corpus, freely available as part of the CHILDES collection of child-directed language corpora (MacWhinney, 2000). Since the sound recordings for the Korman corpus are too faint to be utilized by ASR, the audio-recordings for a subsection of the Brent corpus (Brent & Siskind, 2001), also available through CHILDES/TalkBank (MacWhinney, 2000), were used for Simulation 1.

METHOD

Input data

The input corpus for Simulation 1 was based on recordings of four mothers from the Brent corpus, identified by the codes *c1*, *f1*, *f2* and *q1*, directed at infants age 0;9 to 0;10·26. Since a large proportion of the experimental literature examining word segmentation focuses on infants between 0;7 and 0;11, it has here been assumed that the Christiansen model is best applied to input directed to infants within this time period. Recordings directed at infants older than 0;11 were excluded from this study as being beyond the age most appropriate for the model. Recordings earlier than 0;9 are rare in the Brent corpus and usually record the mother's first recording session. They were excluded to avoid self-conscious speech and other effects of first-time recordings.

Using the transcriptions' CHAT codes, we removed utterances containing any type of input that might cause trouble for the forced-alignment step, including: whispered or sung speech; unintelligible, untranscribed or partial words; word play or pet names; and mentions of the family's last name (left untranscribed to preserve anonymity). This left 13,443 utterances for the four mothers. Using HMM-based acoustic models trained on the TIMIT corpus (Garofolo, Lamel, Fisher, Fiscus, Pallett & Dahlgren, 1993) of read speech, we phonetically aligned this subset of the Brent corpus, performing a forced alignment on the canonical pronunciations found in the CMU dictionary (1993). We utilized the resulting phonetic boundaries to segment each utterance into individual phones. We then calculated the average frame

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

likelihood of the twenty best monophones for each segment and converted these likelihoods into posteriors by normalization.

While in typical ASR tasks it is unusual to first train the models on the same material that will be evaluated, it should be noted that what we are trying to derive is an approximation of the phonetic confusability in the acoustics. Thus, if the models are trained on one phone but during testing they prefer another, this is a clear indication of acoustic confusability, and we can have more confidence that misrecognitions are not due to training/test mismatch.

In order to further increase the confidence in these phonetic materials, utterances that did not have good performance in phone classification across the entire utterance were discarded. The performance of the phone classification across an utterance was calculated using a measure called APPROXIMATE ACCURACY, defined as the number of phones correctly detected within the top two guesses for each phone. Using this definition rather than exact accuracy allows for more of the desired variation while ensuring that the correct phone was a good candidate, suggesting that the automatic phonetic alignment process was valid.

Two subsets of the Brent corpus were created: one that has utterances of approximate accuracy of at least 33·3% and that had more than one phone classified correctly (hereafter called 'Large Brent'), and a second, higher-confidence corpus that had an approximate accuracy of at least 60% per utterance (hereafter called 'Small Brent'). The 60% cut-off point was chosen to represent a subset of utterances with levels of acoustic variation roughly comparable to that found in low-noise speech corpora such as TIMIT: the rate of canonical pronunciations is on the order of 60–80% in TIMIT, depending on stress and syllabic position (Fosler-Lussier, Greenberg & Morgan, 1999). The 33·3% cut-off, on the other hand, allows a more representative sample of the utterances in the Brent corpus, including a number of longer, more complex utterances. Each of these two subsets were further divided 90%—10% into training and test corpora, as shown in Table 1.

Model design

The training procedure was the same as for CAS98, with the exception that the input and target vectors for the PROBABILITY VECTOR condition used are not binary, but continuous (rounded to four decimal places) in the range [0, 1]. While training on probabilistic targets may be unusual, and is a departure from the way uncertainty is handled in CA97, this method of training is consistent with the assumption made here that infants at this stage of development do not have access to the phonemic identity of the target segment, except through the probabilistic cues encoded in the input.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

TABLE 1. *Size of the training and test corpora for the two Brent corpus subsets*

Corpus	Utterances	Word tokens	Word types
The 60%-accuracy subset (Small Brent)			
Training	2861	6443	782
Testing	316	740	258
Total	3177	7183	819
The 33%-accuracy subset (Large Brent)			
Training	7030	22,193	1493
Testing	781	2486	552
Total	7811	24679	2592

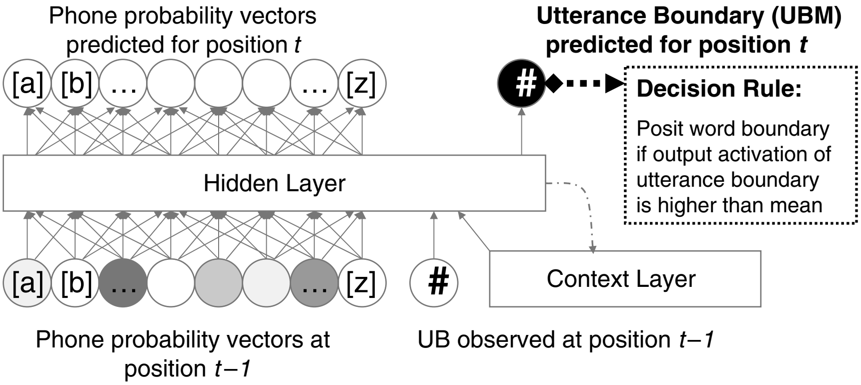
Phone probability vectors predicted for position *t*
Utterance Boundary (UBM) predicted for position *t*
[a]
[b]
…
…
[z]
#
Decision Rule:
Posit word boundary if output activation of utterance boundary is higher than mean
Hidden Layer
[a]
[b]
…
…
[z]
#
Context Layer
Phone probability vectors at position *t–1*
UB observed at position *t–1*

Fig. 2. The phone-probability-vector network used in the probability-vector conditions of Simulation 1. Shades of grey in input units indicate graded activation.

In order to keep the phoneset the same as that used in CAS98, the phone probability vectors produced by the APC system were converted from the 61-phone TIMIT phoneset to the 36-phone MRC phoneset that CAS98 used. The twenty best monophones' posterior probabilities were normalized to sum to 1, and used as input activations for the corresponding input units.[6] A schematic representation of the network used is given in Figure 2.

For comparison purposes, we also provide a CANONICAL (or citation-form) version of both Brent subsets, trained with fully symbolic input taken from the canonical pronunciations of each word as listed in the CMU dictionary, converted to the MRC phoneset (without stress information). This

[6] An anonymous reviewer notes that the normalization of the output and target layers such that the phone probability vectors summed to 1 may have made training more difficult for the SRN, and adversely affected performance for reasons unrelated to the model's ability to handle subsegmental variation. This possibility will need to be evaluated in future work.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

TABLE 2. *Results from Simulation 1a: precision and recall for the 37-70-37 phon-ubm SRN trained and tested with canonical, citation-form input and with automatically phone-classified probabilistic input from the Small Brent corpus subset, compared with two baselines (Prec. = Precision; Rec. = Recall)*

	Boundary		Word		Lexicon	
Input or baseline type	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Canonical	0·589	0·772	0·291	0·381	0·250	0·176
Probability vector	0·577	0·773	0·258	0·345	0·218	0·196
Baselines						
Length-based	0·515	0·613	0·204	0·243	0·133	0·155
Utterance-as-word	1·000	0·427	0·462	0·197	0·250	0·108

corresponds to the PHON-UBM condition in CAS98. Since there is no exact correspondence to stress in the probability vector condition, the PHON-UBM-STRESS condition is not reported for the Brent corpus. In addition, two baselines were used, following CAS98. The first one (the UTTERANCE-AS-WORD baseline) posits word boundaries only at utterance boundaries, treating each utterance as a single word. The other (the LENGTH-BASED baseline) learns from the training corpus the distribution of word lengths (in segments) but gathers no information relating to the identity of the segments. It then chooses word lengths randomly from the distribution of lengths learned.

RESULTS AND DISCUSSION

Simulation 1a: the Small Brent subset

Results are reported first for the smaller, more restrictive subset of the Brent corpus, shown in Table 2. This corpus subset (the Small Brent corpus) has a much shorter mean utterance length than the Korman corpus used in CAS98 or the Large Brent subset used in Simulation 1b and Simulation 2 (2·3 words per utterance, as opposed to 3·2 for Large Brent and 3·1 for Korman), and a much higher incidence of single-word utterances (46% for Small Brent, versus 28% for Large Brent and 26% for Korman). It follows that the single-word baseline will have substantially better recall on the Small Brent subset than on other corpora.

Simulation 1a.i: the canonical condition. The SRN using the canonical transcription performs above the length-based baselines for all measures except type recall ($\chi^2(1) = 10{\cdot}0$, $p = 0{\cdot}0016$ for boundary precision; $p < 0{\cdot}001$ for all other comparisons). As expected, the SRN's performance on boundary precision is trivially worse than the utterance-as-word baseline's perfect precision, and the boundary recall is (trivially) better. However, due

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

to the large number of one-word utterances in the Small Brent subset, the utterance-as-word baseline outperforms the SRN on word precision as well ($\chi^2(1)=30{\cdot}6$, $p<0{\cdot}001$), and matches it on type precision ($\chi^2(1)=0{\cdot}0075$, $p>0{\cdot}9$). Therefore, the performance of the SRN is not as clearly superior to the baselines as it is for CAS98's simulations using the Korman corpus.[7]

Simulation 1a.ii: the probability vector condition. The SRN trained and tested on the probability vector input performs fairly similarly to those using the canonical transcription. Just like the canonical-input SRN, the SRN using the probability vector input also outperforms the length-based baseline for all measures except type recall ($\chi^2(1)=6{\cdot}8$, $p=0{\cdot}009$ for boundary precision; $\chi^2(1)=7{\cdot}0$, $p=0{\cdot}008$ for word token precision; $p<0{\cdot}001$ for all other comparisons). The two SRNs do not differ significantly on any of the six measures. Although the performance on word precision and recall appears to be lower for the probability vector condition, the differences are not statistically significant ($\chi^2(1)=2{\cdot}7$, $p=0{\cdot}10$ for word token precision; $\chi^2(1)=1{\cdot}9$, $p=0{\cdot}16$ for word token recall).[8]

Simulation 1b: the Large Brent subset

Because the relatively small size and short average utterance length of the Small Brent subset made it difficult to distinguish the performance of the SRNs in the canonical and probability vector conditions from the baseline, it is necessary to examine a larger subset of the Brent corpus to obtain reliable figures. It is also useful to see how the Christiansen model (with the types of input examined here) fare with a greater degree of subsegmental variation than that provided in the Small Brent subset. The Large Brent subset, more than double the size of the Small Brent subset, makes this closer look possible. Results for this corpus subset are shown in Table 3.

Simulation 1b.i: the canonical condition. The canonical SRN outperforms the length-based baseline on all measures ($p<0{\cdot}001$ on all comparisons) except type recall ($\chi^2(1)=0{\cdot}26$, $p>0{\cdot}6$). As seen in Simulation 1a.i with the Small Brent corpus, the SRN performs worse than the single-word baseline on word precision ($\chi^2(1)=11{\cdot}0$, $p<0{\cdot}001$), but better than baseline on boundary, word and type recall ($p<0{\cdot}001$ on all comparisons). Unlike

[7] It should be noted that the performance of the canonical model as replicated here is considerably below that reported by Christiansen *et al.* (1998) on the Korman corpus. The CAS98 study reported lexical boundary precision and recall at 0·659 and 0·713 and word (token) segmentation precision and recall at 0·373 and 0·404. The reason for this discrepancy is not known.

[8] The same result is obtained using the distributed input based on the phonological features described in CAS98. When the features from Christiansen *et al.* (2005) are used, the probabilistic-input SRN is significantly worse than the canonical-input SRN in word token precision and recall ($\chi^2(1)=10{\cdot}6$, $p=0{\cdot}0011$ for word token precision; $\chi^2(1)=9{\cdot}8$, $p=0{\cdot}0017$ for word token recall). Other measures are not significantly different.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

TABLE 3. *Results from Simulation 1b: precision and recall for the 37-70-37 phon-ubm SRN trained and tested with canonical, citation-form input and with automatically phone-classified probabilistic input from the Large Brent corpus subset, compared with two baselines (Prec.=Precision; Rec.=Recall)*

	Boundary		Word		Lexicon	
Input or baseline type	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Canonical	0·531	0·861	0·233	0·377	0·226	0·261
Probability vector	0·482	0·720	0·148	0·222	0·131	0·276
Baselines						
Length-based	0·464	0·533	0·150	0·173	0·095	0·276
Utterance-as-word	1·000	0·314	0·288	0·091	0·122	0·129

Simulation 1a.i, the SRN outperformed the single-word baseline on type precision ($\chi^2(1)=22{\cdot}1$, $p<0{\cdot}001$).

Simulation 1b.ii: the probability vector condition. Unlike Simulation 1a, for the Large Brent corpus the subsegmental variation affects performance significantly. The SRN trained and tested on the probability vector input performs significantly worse in all measures compared to the SRN trained and tested on the canonical input ($p<0{\cdot}001$ for all comparisons), except type recall ($\chi^2(1)=0{\cdot}25$, $p>0{\cdot}6$). This drop in performance is sufficient to bring boundary and word precision down to the level of the length-based baseline ($\chi^2(1)=1{\cdot}99$, $p>0{\cdot}1$ for boundary precision; $\chi^2(1)=0{\cdot}03$, $p>0{\cdot}8$ for word precision). Still, boundary recall ($\chi^2(1)=184{\cdot}9$, $p<0{\cdot}001$) and word recall ($\chi^2(1)=18{\cdot}6$, $p<0{\cdot}001$) are significantly better than the length-based baseline, as is type precision ($\chi^2(1)=8{\cdot}71$, $p=0{\cdot}0031$).

DISCUSSION

Simulation 1a suggests that the Christiansen model, even without the stress cue, is robust to data with subsegmental variation when this variation is carefully controlled. This finding is consistent with previous tests of the Christiansen model in CA97. Because the near-continuous vector output of a phone recognition classifier is a more accurate representation of human perception than the feature byte-swapping done in CA97, this study provides added support for the model's basic robustness. The corpus used in this simulation is small and the language used very simple, so the claim of success has to be qualified.

Simulation 1b, performed on a larger subset of the Brent corpus, including utterances that are considerably more difficult both for the phone classifier and for the Christiansen word segmenter, shows that there is a point where the variation does cause significant degradation to the model's performance.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

The best interpretation for this degradation is not immediately clear. One possible explanation is that large-scale variation compromises the reliability of the segmental cues, such that it is no longer possible to find word boundaries using these cues alone.[9] Christiansen's network was explicitly designed to combine multiple cues in a plausible way, without direct supervision of the word segmentation task itself. Therefore, when moving to spoken language, it is worth looking for a set of more robust cues which can, in combination with the segmental cues, improve the model's word segmentation performance. This would be a direct validation of Christiansen's idea, and all the more compelling because of the use of more naturalistic input.

SIMULATION 2

To further study the degradation of performance observed in Simulation 1b, we introduce other cues. We do this while continuing to require that the cues that are used are plausibly available to an infant aged 0;8.

The lexical stress cue used by CAS98 was derived from a dictionary, so it cannot be assumed to be available directly. However, there is ample evidence that infants of the appropriate age use many of the acoustic cues associated with lexical stress, such as pitch, duration and spectral tilt (Thiessen & Saffran, 2004). It is also likely that they are able to distinguish between degrees of care in the articulation of a syllable or segment.

We measure correlates of articulatory care, estimating them from the phone probability vectors already available in the APC system. Our use of these cues is motivated in part by an assumption that babies benefit most from stretches of speech that they can readily interpret. This is an extrapolation of findings that infants prefer speech over non-speech, and CDS over adult-directed speech (e.g. Fernald, 1985). This cue may also be interpreted as a proxy measure of 'local hyperarticulation' (Cho & Keating, 2007), associated in English both with lexical stress (e.g. de Jong, 1995) and the beginnings of words (e.g. Fougeron & Keating, 1997).

Finding the start of these salient, more easily interpretable stretches may facilitate word learning. Simulation 2 therefore incorporates an additional cue to signal the onset of an acoustically distinct stretch of speech, or region of local hyperarticulation.

We assume here that the confusion matrix of the automatic phone classifier approximates the perceptual confusions of an infant learner of the language. If the APC assigns a high posterior probability to just one phone, and much lower probabilities to all the others, we assume that this reflects

[9] Another possibility is that the normalization of the output and target layers caused difficulties in training (as noted in footnote 6). However, if that is the case, it is not clear why it affected the network only in the Large Brent corpus and not the Small Brent corpus subset.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

a careful and clear pronunciation of that phone. Conversely, if the phone classifier's activation is spread nearly equally between a large number of possible phones, we treat this as evidence that the phone is unclear and/or sloppily articulated. Our ideal measure, which is a measure of posterior entropy, will be referred to as SEGMENTAL CONFIDENCE; it is approximated here by the maximum activation value output by the phone classifier, regardless of the identity associated with that value.

METHOD

Input data

The input data used in Simulation 2 is the same as that used in the probability vector condition of Simulation 1 (i.e. Simulations 1a.ii and 1b.ii), using the same two corpus subsets (Small Brent and Large Brent) as above. As the patterns observed in the two corpus subsets are similar, and the Large Brent corpus allows differences between the two conditions to be seen more clearly, only the results for the Large Brent corpus are reported here.

Model design

In Simulation 1, the activation of each input unit is determined by the posterior probability of each phone given the acoustic signal for a particular time interval of the speech signal (corresponding to the duration of a single phone in the forced alignment). In Simulation 2, we use these same input units, but augment them by adding input units corresponding to the degree of perceived clarity of the input. We operationalize the model's confidence in the input by taking the maximum probability value in the segment's phone probability vector, including it as an additional cue. The segmental confidence for segment position *t* ($\mathrm{SC_t}$) may be written as in Equation 3, where A is the phoneset, X_t is the acoustic information at position *t*, and $\Pr[Q_t = q \mid X_t]$ is the probability of the phone q occurring at timestep *t*.

$$SC_t = \max_{q \in A} \Pr\left[Q_t = q | X_t\right] \tag{3}$$

Just as stretches of phonemically distinct speech are found by measuring areas of high segmental confidence, the beginnings of such stretches can be found by noting segments where the segmental confidence is larger than that of the preceding segment. The delta segmental confidence for segment position *t* ($\Delta\mathrm{SC_t}$) is approximated here with the following function:

$$\Delta SC_t = \begin{cases} SC_t - SC_{t-1}: & SC_t > SC_{t-1} \\ 0: & SC_t \leq SC_{t-1} \end{cases} \tag{4}$$

For this simulation, we assume that both absolute segmental confidence for a given segment and the amount of increase in segmental confidence

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

from the previous segment are equally important for detecting the starting points of potential islands of reliability. Accordingly, we add two additional input and output units corresponding to $\mathrm{SC_t}$ and $\Delta\mathrm{SC_t}$, and remove sufficient hidden and context units to keep the number of weighted connections as close to constant as possible.

In this simulation, as in Simulation 1, two distributed, feature-based representations of the input were tested, corresponding to the representations used in CAS98 and Christiansen *et al.* (2005), in addition to a localist representation. As in Simulation 1, the two feature-based input representations show the same trends as the localist representation, and do not perform significantly better; hence, nothing of interest is gained in reporting them. Therefore, only the performance of the localist representation is reported here.

Evaluation procedure

Simulation 2 seeks to answer two questions: first, whether an automatically derived non-segmental cue, such as segmental confidence as defined above, can serve as a useful 'catalyst task' analogous to the way the original Christiansen model used lexical stress; and second, whether the information contained in the segmental confidence cue is helpful for finding word beginnings when combined directly with the Christiansen model's utterance boundary prediction task. Accordingly, the evaluation for Simulation 2 is conducted twice, as two separate sub-experiments.

The first evaluation procedure (Simulation 2a) treats segmental confidence and ΔSC simply as additional catalyst features in the input and output levels of the Christiansen network, just as lexical stress was treated in CAS98. The method for positing word boundaries is the same as the original Christiansen model and Simulation 1, depending only on the UBM output unit's activation at the previous segment – that is, a boundary is posited between segments $(t-1)$ and t only if $actv(UBM)_{t-1}$ is greater than the mean activation for the UBM output unit. This evaluation procedure intuitively measures the effect of the segmental confidence cues in the network's ability to generalize from ends of utterances to ends of words.

The other evaluation (Simulation 2b) uses a new method of combining the network's prediction of an utterance boundary at segment position t (i.e. $actv(UBM)_{t-1}$) with the network's judgment of whether or not the segment at position t begins a region of clear speech. We name this second criterion SEGMENTAL SALIENCE (SS) and define it as the sum of absolute and delta segmental confidence: $SS_t = SC_t + \Delta SC_t$. The full decision rule compares the product of the UBM activation at the previous segment and segmental salience at the current segment to a threshold. Analogous to Simulation 1, the threshold is set equal to the mean value for this product at each point

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

over the entire corpus, as shown in Equation 5. This evaluation method examines the effect of the segmental confidence cues (combined as single segmental salience cue) on predicting word boundaries directly, without presupposing their effect on generalizations from utterance boundaries. A schematic illustration of the modified Christiansen network with this decision rule is given in Figure 3.

$$Boundary(t-1, t) = \begin{cases} 1: \ actv(UBM)_{t-1} * SS_t > \ \overline{actv(UBM)_{t-1} * SS_t} \\ 0: \ otherwise \end{cases} \qquad (5)$$

RESULTS AND DISCUSSION

Simulation 2a: segmental confidence as an additional catalyst task

Simulation 2a examines the performance of the SRN trained and tested on the probability vector input of the Large Brent corpus subset with the two segmental confidence cues (absolute and delta) used only as additional immediate tasks for the network to solve. This scenario is directly analogous with the phon-ubm-stress condition in CAS98. However, unlike CAS98, the extra cue yields no improvement in performance. The SRN using the extra cues performs no better than the SRN trained without segmental confidence, shown in Simulation 1b.ii. Although it shows a slight improvement in terms of boundary recall, it performs worse in terms of boundary precision ($\chi^2(1) = 4{\cdot}41$, $p = 0{\cdot}0358$) and word token precision ($\chi^2(1) = 4{\cdot}52$, $p = 0{\cdot}0335$), and no better in word token recall, type precision or type recall.

Simulation 2b: segmental salience as an additional criterion

When the segmental salience cue is used as an additional CRITERION for placing word boundaries, it does improve performance. The performance of the ‘segmental salience as criterion’ condition outperforms the ‘segmental confidence as catalyst’ condition tested in Simulation 2a. While performance in boundary recall is worse than in Simulation 2a, the segmental salience cue improves performance in boundary precision ($\chi^2(1) = 46{\cdot}43$, $p < 0{\cdot}001$) and word token precision ($\chi^2(1) = 78{\cdot}09$, $p < 0{\cdot}001$), and in word token recall ($\chi^2(1) = 10{\cdot}68$, $p = 0{\cdot}0011$) and type recall ($\chi^2(1) = 10{\cdot}64$, $p = 0{\cdot}0011$).

Analogous differences are observed when comparing the results of Simulation 2b to the recognized input without any additional cues (in Simulation 1b.ii). While performance in boundary recall for the ‘segmental salience as criterion’ condition is worse than the probability vector condition in Simulation 1b.ii, performance in boundary and word token precision is better ($p < 0{\cdot}001$ for all comparisons). Word token recall also improves ($\chi^2(1) = 6{\cdot}55$, $p = 0{\cdot}0105$), as does type recall ($\chi^2(1) = 5{\cdot}53$, $p = 0{\cdot}0187$).

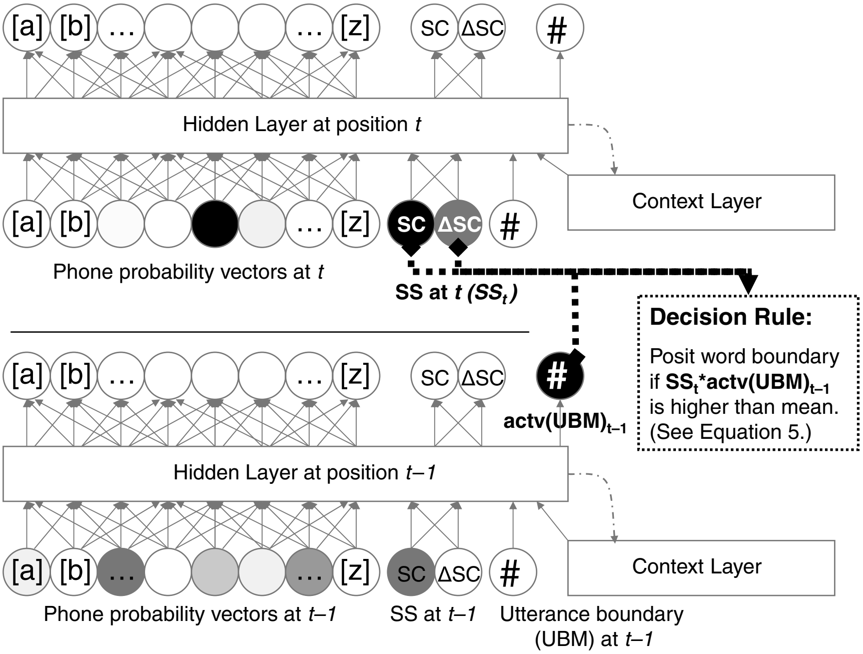
[a]
[b]
…
[z]
SC
ΔSC
#
Hidden Layer at position t
Context Layer
Phone probability vectors at t
SS at t (SS_t)
Decision Rule:
Posit word boundary if SS_t*actv(UBM)_t–1 is higher than mean. (See Equation 5.)
actv(UBM)_t–1
Hidden Layer at position t–1
Context Layer
Phone probability vectors at t–1
SS at t–1
Utterance boundary (UBM) at t–1

Fig. 3. The 39-68-39 phon-ubm-SC network for two subsequent time-steps, $t-1$ (bottom) and t (top), for the 'segmental salience as criterion' condition tested in Simulation 2b.

Performance is in some measures comparable to the canonical phon-ubm condition in Simulation 1b.i. While performance in boundary and word recall is worse than in Simulation 1b.i, performance is better in type recall ($\chi^2(1)=8{\cdot}49$, $p=0{\cdot}0036$). Indeed, the 'segmental salience as criterion' condition shows a stronger performance in type recall than any other variant of the Christiansen model tested on the Brent corpus. Results for Simulation 2 (both 2a and 2b) are given in Table 4.

DISCUSSION

Simulation 2 seeks to test two claims of CAS98. First, and more generally, it tests the claim that an algorithm that combines multiple cues to word segmentation performs better than any one cue alone. Second, and more specifically, it tests whether a single Elman network can effectively combine multiple cues via simultaneous training on multiple prediction tasks (or 'catalyst' tasks), without direct supervision on the target task (word segmentation).

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

TABLE 4. *Results from Simulation 2: precision and recall for the* 39-68-39 *phon-ubm-SC SRN trained and tested with automatically phone-classified probabilistic input from the Large Brent corpus subset, aided by segmental confidence information added and evaluated* (*i*) *as an extra catalyst task and* (*ii*) *as an extra word boundary placement criterion, compared with two baselines* (*Prec.*=*Precision*; *Rec.*=*Recall*)

	Boundary		Word		Lexicon	
Input or baseline type	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Segmental confidence as catalyst	0·458	0·744	0·132	0·213	0·123	0·252
Segmental salience as criterion	0·541	0·648	0·211	0·253	0·137	0·343
Baselines						
Length-based	0·464	0·533	0·150	0·173	0·095	0·276
Utterance-as-word	1·000	0·314	0·288	0·091	0·122	0·129

A replication of CAS98 (not presented here) found evidence for both of these claims, when the training and testing corpora used symbolic transcriptions without subsegmental variation, and dictionary-derived stress cues rather than hyperarticulation cues. Simulation 2, which incorporates larger amounts of variation than Christiansen reports testing, also finds evidence for the first, more general claim. When Christiansen's word segmentation model is faced with highly variable, potentially ambiguous input, multiple probabilistic cues still outperform each cue separately – when the evaluation procedures consider all the cues. When segmental salience is combined with the activation of the utterance boundary marker as part of the final decision rule for positing word boundaries (as in Simulation 2b), performance improves over the use of the recognized segmental information alone.

However, when segmental confidence and delta segmental confidence are used as catalyst features for predicting utterance boundaries only, and utterance boundary prediction alone is used to posit word boundaries, performance is no better than when using recognized segmental information alone. This was a surprising result, and not fully understood. Possible implications of this finding will be discussed further in the next section.

One interesting finding in Simulation 2b is the exceptionally high type recall – the fraction of distinct word types in the corpus that were correctly segmented – in the 'segmental salience as criterion' condition, compared with other conditions tested on the Large Brent corpus subset. This is appropriately analogous to the high performance of the phon-ubm-stress condition in the original CAS98 study, suggesting that the segmental salience cue may make similar contributions in this condition as the stress cue did in CAS98. While space does not allow for a detailed study of this finding, it

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

appears that the segmental salience cue is acting similarly to lexical stress, in that it correctly segments stressed, word-initial syllables from the preceding word, as in 'what#cha#doing' and 'are#you#fish(ing)', where other variants of the model miss the underlined word boundary. The SS cue likewise prevents the placement of word boundaries before non-word-initial syllables without primary stress, as in 'fish(ing)' and 'peeka(boo)', where other model variants split these up. It also helps within syllable boundaries, keeping together complex codas such as 'than(k)', 'plan(t)', 'blin(d)' and 'tir(ed)', where other variants split off the final stop. Since these stops are known to have different allophones word-initially vs. word-finally, it seems plausible that the 'segmental salience as criterion' variant, aided by the probabilistic input, is learning and using this allophonic information to place the word boundary correctly in these examples.

GENERAL DISCUSSION AND CONCLUSIONS

Despite the growing number of high-quality audio corpora available, such as the Brent corpus, most models of the word segmentation task represent an utterance as a sequence of symbols. For older models, this idealization was no doubt necessary given the available resources at the time. Nonetheless, it is problematic. It is doubtful that babies perceive phonemes with complete reliability and accuracy – or even that adults are completely reliable and accurate at perceiving phonemes independently of 'top-down' information from higher levels. Certainly, infants can distinguish between phonemes in their soon-to-be-native language in certain conditions, while successfully ignoring or disregarding distinctions that are non-phonemic in their language (Werker & Tees, 1984) as early as 0;10. Still, it does not follow that babies are able to apply this ability to all tasks or situations.

Even if infants are able to apply their phonemic knowledge to the word segmentation task, it does not follow that they do so with 100% accuracy, as mentioned above. Polka & Rvachew (2005) report a discrimination accuracy of 80% for healthy infants between 0;6 and 0;9 on a simple, two-alternative choice between /bu/ and /gu/. In running speech, infants' accuracy is probably lower still, at least until the infants learn enough contextual information to begin applying top-down processing. Indeed, some phonemic pairs are indistinguishable until relatively late: babies natively learning Tagalog do not even perform above chance in distinguishing /n/ from /ŋ/ in syllable onsets until 0;10, even though in Tagalog /n/ and /ŋ/ are phonemically distinct in syllable onsets (Narayan, Werker & Beddor, in press).[10]

[10] The Christiansen model, like most of the computational models of word segmentation reviewed, does not commit explicitly to a precise age range of infants being modeled. Christiansen *et al.* (1998: 253) suggest that they are focusing on initial stages of word segmentation, and note that their speech data was directed at infants at 6–16 weeks of

Even for phonetically trained adults, 100% accuracy of phone identification without the benefit of higher-level cues seems unrealistic. The Buckeye corpus, which contains spontaneous adult-directed speech, reports 80·3% overall agreement (*kappa*=0·797) with unanimous agreement on only 62% of the segments (Pitt *et al.*, 2005). In contrast, models that use phonemic transcriptions of words implicitly assume that infants are receiving, with 100% accuracy and confidence, exactly what the majority of transcribers heard – or (in the case of word-level transcriptions) what the pronunciation dictionary dictates.

If the assumption of 100% accuracy on phonemic discrimination or identification is unrealistic, then it follows that one cannot tell with confidence which computational approaches to word segmentation are most promising based solely on their performance on idealized, symbolic input. The simulations reported here, particularly Simulation 1b, demonstrate that the Christiansen model can perform quite differently on canonical transcription-derived input than on probabilistic, audio-derived input. Since most recent word segmentation models have not been extended to model the latter type of input, the effects of the acoustic variation and ambiguity found in natural speech on the performance of these models are still unknown.

One way to move beyond a reliance on transcriptions alone is to use automatic speech recognition. While this is not a novel idea, previous work has used corpora that either were not widely available (e.g. Roy & Pentland, 2002) or not child-directed speech (e.g. de Marcken, 1996). With the Brent speech corpus, it is now possible for the community to compare on a common corpus with speech input. It is hoped that many models will be re-examined with these or similar data as the Christiansen model has been here.[11]

Testing the Christiansen model with ASR-based input

Two basic claims of CAS98 are (1) that hidden structure can be learned implicitly by training a model on an immediate task (such as segmental prediction) on observable cues, and (2) that the combination of several such cues allows for better learning of hidden structure (such as word boundaries) than any single cue. Simulation 1 re-examines the first of these claims,

age – much younger than the age noted here. Even if a considerably later timeframe of 0;7–0;11 were adopted as relevant for models of word segmentation, corresponding to the stages of word segmentation examined in Christiansen *et al.* (1998), abilities shown at age 0;10 can be assumed to be in place and available only for the final stages of an infant's development of word segmentation abilities.

[11] The authors are happy to make the input used in these studies available to other researchers upon request.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

taking into account the fact that even so-called 'observable' cues are also probabilistic and potentially ambiguous. It re-evaluates the Christiansen model as described in CAS98 with input data derived from the Brent audio corpus via 'soft-decision' automatic phone classification. Simulation 1a suggests that, when the degree of variability in the audio signal is kept within certain bounds, the CAS98 model is robust to this subsegmental variation. However, in Simulation 1b, using a larger and more variable corpus, performance was significantly worse in the recognized condition than the canonical (or citation-form) condition.

These results cast some doubt on the robustness of the Christensen model's ability to segment ambiguous input on the basis of segmental cues alone. Several alternative interpretations of this result are possible. Perhaps the most obvious is that segmental information is inherently less reliable than previous word segmentation models (based largely on segmental information) have suggested, and hence the need to make greater use of other (e.g. suprasegmental and/or subsegmental) cues is greater than supposed. This is consistent with research that suggests that non-segmental cues trump segmental cues for infants aged 0;8 when they conflict (Johnson & Jusczyk, 2001). It is also consistent with recent findings in the context of a closely related problem, sentence boundary detection in speech. Liu (2004) finds that lexical methods of sentence boundary detection are less robust in highly variable speech than prosodic methods, because the high word error rate in the ASR obscures the necessary lexical cues. By analogy, if errorful phone recognition (whether by infant or machine) obscures the cues most needed for word segmentation, then cues unaffected by the phones (such as prosodic cues) will naturally be more robust. If this is the case, then it follows that other word segmentation models that rely exclusively or primarily on segmental cues should perform significantly worse on the Large Brent phone probability vectors than on the canonical data with which they have been tested heretofore. We leave it as an open challenge to the computational modeling community to test this hypothesis with their favorite segment-based word segmentation models.

It follows from this hypothesis that combining other, non-segmental cues should help ameliorate performance. This is CAS98's second major claim, and like the first claim, it was tested only with idealized, symbolic input – word stress as abstracted from dictionary pronunciation guides, not as observed in the actual speech signal. Simulation 2 re-examines this claim, testing the contribution of a non-segmental cue loosely analogous to word stress, but derivable from the audio signal – segmental salience. In a condition analogous to CAS98's phon-ubm-stress condition, this simulation combines probabilistic segmental information together with a measure of confidence in that information, corresponding to the local hyperarticulation found in word-initial positions in English. It also shows that multiple cues outperform

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

single cues, even when those cues are probabilistic – and derived from the same source.

However, it must be noted that Simulation 2 only shows that multiple cues improve performance when they are combined directly as separate criteria in the decision rule, not when merely combined as multiple catalyst tasks in an SRN. This could mean that that there are limits to CAS98's method of combining cues, or that the evaluation measures that CAS98 happened to use are less appropriate in this case. More specifically, using 'catalyst' tasks or features may be limited by the effectiveness of those catalyst features in improving the performance of the task measured during the evaluation – in the case of CAS98, the extrapolation from utterance boundaries to word boundaries. It appears from Simulation 2a that the segmental confidence catalyst features, while possibly helpful in identifying word boundaries in some other way, were not learned effectively by the network and/or did not contribute to a better generalization from utterance boundaries to word boundaries. This does not disprove the usefulness of these cues to the word segmentation task nor invalidate the Christiansen model per se, but it does suggest the limitation of evaluating its performance based on a single variable such as the UBM unit output.

In this case, a more direct combination of heuristics seems more appropriate and effective than treating all cues as prediction tasks for a single heuristic. The results of Simulation 2b suggest that it makes more sense to treat the segmental salience feature as a direct cue to word BEGINNINGS, rather than as an additional feature for predicting which utterance endings extrapolate well to word endings.[12]

Implications and limitations of the findings

The 'segmental salience as criterion' condition in Simulation 2b outperforms the corresponding 'segmental confidence as catalyst' condition in Simulation 2a. This suggests a segmentation strategy that infants might be using. Perhaps they are able to detect regions of clear speech, and treat the beginnings of such regions as likely word boundaries. This possibility will need to be tested experimentally in humans.

We also need to know more about the corpus used. The types of variation found in the Small Brent subset clearly correspond to those expected from normal variation speech such as reduced or casual pronunciations of words, allophony and dialect differences. By contrast, the Large Brent subset

[12] The same is probably true of lexical stress as well – had CAS98 evaluated lexical stress as a separate contributor to their method for positing word boundaries during their evaluation, they likely would have had even larger precision and recall scores.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

includes a much higher proportion of ASR errors, and many of these have no easily recognizable linguistic explanation. In ASR, such patterns of error are often due to extraneous factors such as background noise. It would be worth checking whether these utterances are sufficiently messy and noisy to give human listeners trouble. It is not obvious how best to do this. This is an issue which is always likely to arise with simulations of development: the models inevitably have properties that go beyond what is known about the target behavior, and, absent further study of the actual process of human development, it is unclear whether these properties are desirable. We cannot safely commit to the Large Brent corpus until we know to what extent it represents the challenges that a learner faces, and we can only know that if we know enough about these challenges in the first place. It is known that adult humans are excellent at tolerating noise levels that would leave current ASR systems floundering, but unknown whether infants also share this capacity, or, if they do not initially have it, how they acquire it, and with what timecourse.

Future directions

The simulations described here confirm the general claims of CAS98. They also point out areas for further investigation and refinement of the model. Cue integration is an especially rich opportunity. Because each cue will be error prone to a different degree and in a different way, Christiansen and Allen's account needs to be augmented with details of how this happens and how the cognitive system manages this systematic uncertainty. This problem is a qualitatively different extension of the original. We have argued above that other approaches to word segmentation should be evaluated primarily by testing their performance on input that preserves the variation naturally present in speech. Based on our results there is no strong reason to suppose that strategies and learning biases that work for idealized symbolic input will have the same properties in noisier and messier environments.

There is great need for deeper understanding of language acquisition in adverse contexts, including noisy environments (cf. Newman, 2005), temporary hearing loss (cf., e.g., Polka & Rvachew, 2005), and profound hearing loss and/or cochlear implants. Using ASR models with different types of added noise, or even with input derived from a cochlear implant's output, could provide models for how children might segment and acquire language in adverse conditions. A beneficial side effect of this enterprise is to motivate a more varied set of evaluation techniques for automatic methods of processing speech. Not only can psychology benefit from ASR technology, it can also give back suggestions for how to do evaluations that are better focused on the task than traditional measures of word error rate.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

REFERENCES

Aslin, R. N., Woodward, J. Z., LaMendola, N. P. & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In K. Demuth & J. L. Morgan (eds), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, 117–34. Mahwah, NJ: Lawrence Erlbaum Associates.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition* **83**, 167–206.

Blank, D., Kumar, D., Meeden, L. & Yanco, H. (2003). Pyro: A Python-based versatile programming environment for teaching robotics. *Journal of Educational Resources in Computing* **3**, 1–15.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning* **34**, 71–105.

Brent, M. R. & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition* **81**, 31–44.

Cairns, P., Shillcock, R., Chater, N. & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus based approach to speech segmentation. *Cognitive Psychology* **33**, 111–53.

Carterette, E. C. & Jones, M. H. (1974). *Informal speech: Alphabetic and phonemic texts with statistical analyses and tables*. Berkeley, CA: University of California Press.

Cho, T. & Keating, P. A. (2007). Effects of initial position versus prominence in English. *UCLA Working Papers in Phonetics* **106**, 1–33.

Christiansen, M. H. & Allen, J. (1997). Coping with variation in speech segmentation. In A. Sorace, C. Heycock & R. Shillcock (eds), *Proceedings of the GALA '97 conference on language acquisition: Knowledge representation and processing*, 327–32. Edinburgh: Edinburgh University Press.

Christiansen, M. H., Allen, J. & Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes* **13**(2/3), 221–68.

Christiansen, M. H., Conway, C. M. & Curtin, S. (2005). Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J. W. Minett & W. S. Wang (eds), *Language acquisition, change and emergence: Essays in evolutionary linguistics*, 205–249. Hong Kong: City University of Hong Kong Press.

CMU (1993). *The Carnegie Mellon pronouncing dictionary, version 0.6*. Pittsburgh, PA: Carnegie Mellon University. Retrieved from www.speech.cs.cmu.edu/cgi-bin/cmudict.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development* **8**, 181–95.

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. In *Proceedings of the 46th annual meeting of the Association for Computational Linguistics: Human language technologies*, 130–38. Presented at the ACL'08, Columbus, OH: ACL.

Fosler-Lussier, E., Greenberg, S. & Morgan, N. (1999). Incorporating contextual phonetics into automatic speech recognition. In John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds), *Proceedings of the International Congress of Phonetic Sciences*, 611–14, San Francisco. Berkeley, CA: University of California, Berkeley.

Fougeron, C. & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America* **101**, 3728–40.

Frank, M. C., Goldwater, S., Mansinghka, V., Griffiths, T. L. & Tenenbaum, J. (2007). Modeling human performance in statistical word segmentation. In D. S. McNamara & J. G. Trafton (eds), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, 281–86. Austin, TX: Cognitive Science Society.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. & Dahlgren, N. L. (1993). *DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM*. Available from www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

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

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press

Halberstadt, A. K. & Glass, J. R. (1997). Heterogeneous acoustic measurements for phonetic classification. In *Proceedings of Eurospeech '97*, 401–404. Rhodes: European Speech Communication Association.

Johnson, E. K. & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language* **44**(4), 548–67.

de Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America* **91**, 491–504.

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language* **5**, 44–45.

Krull, D. (1990). Relating acoustic properties to perceptual responses: A study of Swedish voiced stops. *The Journal of the Acoustical Society of America* **88**, 2557–70.

Liu, Y. (2004). Structural event detection for rich transcription of speech. Unpublished doctoral dissertation, Purdue University.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

de Marcken, C. G. (1996). Unsupervised language acquisition. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

McMurray, B. & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition* **95**(2), B15–B26.

Narayan, C. R., Werker, J. F. & Beddor, P. S. (in press). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*.

Newman, R. S. (2005). The cocktail party effect in infants revisited: Listening to one's name in noise. *Developmental Psychology* **41**, 352–62.

Newman, R. S., Bernstein Ratner, N., Jusczyk, A. M., Jusczyk, P. W. & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology* **42**, 643–55.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S. & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* **45**, 89–95.

Polka, L. & Rvachew, S. (2005). The impact of otitis media with effusion on infant phonetic perception. *Infancy* **8**, 101–117.

Redford, M. A. & Diehl, R. L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *The Journal of the Acoustical Society of America* **106**, 1555.

Roy, D. & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science* **26**(1), 113–46.

Rytting, C. A. (2007). Preserving Subsegmental Variation in Modeling Word Segmentation, or The Raising of Baby Mondegreen. Unpublished doctoral dissertation, The Ohio State University.

Scharenborg, O., Norris, D., ten Bosch, L. & McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science* **29**(6), 867–918.

Thiessen, E. D. & Saffran, J. R. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception and Psychophysics* **66**, 779–91.

Werker, J. & 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.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., *et al*. (2002). *The HTK Book*. Cambridge: Cambridge University Engineering Department.

https://doi.org/10.1017/S0305000910000085 Published online by Cambridge University Press