Learnability of prosodic boundaries: Is infant-directed speech easier?

Bogdan Ludusan[a)] and Alejandrina Cristia
Laboratoire de Sciences Cognitives et Psycholinguistique
EHESS / École Normale Supérieure, PSL Research University / CNRS, Paris, France

Andrew Martin and Reiko Mazuka
Laboratory for Language Development
RIKEN Brain Science Institute, Wako-shi, Japan

Emmanuel Dupoux
Laboratoire de Sciences Cognitives et Psycholinguistique
EHESS / École Normale Supérieure, PSL Research University / CNRS, Paris, France

This study explores the long-standing hypothesis that the acoustic cues to prosodic boundaries in infant-directed speech (IDS) make those boundaries easier to learn than those in adult-directed speech (ADS). Three cues (pause duration, nucleus duration and pitch change) were investigated, by means of a systematic review of the literature, statistical analyses of a new corpus, and machine learning experiments. The review of previous work revealed that the effect of register on boundary cues is less well established than previously thought, and that results often vary across studies for certain cues. Statistical analyses run on a large database of mother-child and mother-interviewer interactions showed that the duration of a pause and the duration of the syllable nucleus preceding the boundary are two cues which are enhanced in IDS, while f0 change is actually degraded in IDS. Supervised and unsupervised machine learning techniques applied to these acoustic cues revealed that IDS boundaries were consistently better classified than ADS ones, regardless of the learning method used. The role of the cues examined in this study and the importance of these findings in the more general context of early linguistic structure acquisition is discussed.

PACS numbers: 43.71.Ft, 43.70.Fq

I. INTRODUCTION

During the first year of life, infants begin learning their native language by acquiring linguistic structure simultaneously at multiple levels (for an overview, see Jusczyk, 2000). How do infants achieve this feat? One possibility is that parents make this task easier by interacting with their infants using a special speech register, a possibility we will call the *hyperspeech hypothesis* (Fernald, 2000). Such a register, characterized by short sentences, repeated words, and exaggerated intonation, has been called Baby Talk, Motherese, Parentese, or, more generally, infant-directed speech (IDS) (Ferguson, 1964). Infants prefer listening to speech produced in the IDS register compared to speech in the unmarked register used with adults (adult-directed speech or ADS) (see Cristia, 2013, for a review of IDS-ADS differences), and the amount of IDS that they are exposed to predicts later language performance (Weisleder and Fernald, 2013; Shneidman and Goldin-Meadow, 2012). It could be that IDS enhances learning purely because of its emotional and social content (Singh *et al.*, 2002). However, the hyperspeech hypothesis makes a stronger claim: that in addition to these effects, the particular linguistic structures present in IDS *simplify* the learning task itself.

Much attention has been given in the acquisition literature to the effects of IDS on the learning of phonetic categories (Kuhl *et al.*, 1997): It has been claimed that in IDS, the vowel space is expanded compared to ADS. Even though subsequent work has substantiated the expansion of the vowel triangle (see Cristia, 2013 for a review), it has also been documented that, in IDS, parents produce vowels more variably (Kirchhoff and Schimmel, 2005), resulting in a net *detrimental* effect of IDS for the learnability of phonetic categories when tested on a large corpus (Martin *et al.*, 2015). The increase in phonetic variability in IDS would seem to contradict the hyperspeech hypothesis at the segmental level, but the hypothesis may still hold true for other linguistic levels. Indeed, there are strong reasons to believe that in the area of prosody, especially that of prosodic boundary markers, IDS could simplify content for the benefit of the language learner.

Linguistic evidence suggests that continuous speech is grouped according to a hierarchy of "suprasegmental" units, from small units (morae, syllables) to several levels of multi-word units (phonological phrases, intonational phrases, and utterances; Nespor and Vogel, 2007). It has been proposed that (large) prosodic units play a fundamental role in early language acquisition, helping with the discovery of words (Johnson *et al.*, 2014) and facilitating syntactic bootstrapping (Morgan and Demuth, 2014).

Learning how to identify these prosodic units is therefore a critical part of the early stages of language acquisition. The boundaries between such units are typically signalled by a combination of acoustic cues (e.g. pause duration, nucleus duration, pitch change) which infants could use to begin segmenting speech.[1] This represents the first step in learning how their language assigns prosodic structure to the speech signal, and it is thus important to understand how these cues are man-

[a)] Electronic address: bogdan.ludusan@ens.fr

ifested in the IDS register, which comprises the bulk of infants' input.

Because IDS is characterized as containing shorter utterances and more exaggerated prosody than ADS (Cristia, 2013), one might expect that prosodic boundaries should be more numerous and easier to find in IDS than ADS. However, it is important not only to document whether or not cues are *exaggerated*, but also to determine whether or not the net result will be enhanced *learnability*. That is the aim of this paper.

The paper is structured as follows: In Section II we review previous analyses of cues to prosodic boundaries in IDS and ADS. In Section III, we present new acoustic measurements using a large corpus containing both IDS and ADS registers. Finally, in Section IV, we directly test the hyperspeech hypothesis through the use of supervised and unsupervised machine learning algorithms applied to the corpus.

II. SYSTEMATIC REVIEW OF PREVIOUS LITERATURE

Phonetic studies in a variety of languages have highlighted the fact that prosodic boundaries may be signalled by at least three types of acoustic cues: pause duration, nucleus duration, and pitch change (Pacia-Cooper and Cooper, 1980). Pause refers to the presence of a silence. However, the perception of a break by adult listeners may integrate duration of both sounds and pauses (Scott, 1982). Nucleus duration refers to the phenomenon of preboundary lengthening by which a given syllabic nucleus (often a vowel) tends to be longer just prior to a boundary compared to a phrase-medial position. Finally, pitch may signal both the cohesion of a stretch of speech as a single prosodic unit and the presence and strength of a boundary in a number of ways, including through a reset in the fundamental frequency (*f0*) level across a boundary.[2]

The goal of the present review is to assess the plausibility of the hyperspeech hypothesis with regard to prosodic boundaries, i.e. the claim that in IDS, prosodic boundaries are more exaggerated or clearer than in ADS. We survey previous studies analyzing the three abovementioned cues: pause duration, nucleus duration and f0 change, and report the results in terms of differences between registers.

We conducted a systematic review, considering all relevant studies on this research topic that fit pre-determined eligibility criteria and summarizing them in a standardized manner. In order to be valid, systematic reviews must be constructed according to precise guidelines; we follow here the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (Moher *et al.*, 2009). It consists of a standardized set of items to be reported, as well as a work flow to be followed. The work flow includes details on the identification, the screening, the eligibility and the decision for inclusion of the considered studies. More detailed information on this methodology, including numerous descriptive tables, can be found in Section S1 of the Supplementary Material.

We chose to conduct both a qualitative and a quantitative review. The sample of studies retained for the *qualitative review* consisted of 19 journal articles, theses or conference proceedings, which contained information on any or all of the three acoustic cues that can signal prosodic boundaries. Of these, 13 were included in the *quantitative review* because they additionally met all of the following criteria: (a) used a corpus of speech collected in the presence of a real child, (b) determined the presence of a boundary using prosodic cues or sentence edges, (c) contained information on both IDS and ADS, and (d) at least one of our two dependent measures (average or median values; and/or inferential statistics) could be found.

When performing the quantitative review, we observed variation in terms of how means and standard errors/deviations were estimated, and how statistical analyses were carried out, since some studies used talkers as units of analysis, and others used tokens (e.g. individual vowel durations). Since it is not meaningful to attempt to derive effect sizes from such heterogeneous studies, we extracted three types of information. First, we noted key methodological information. Second, we extracted mean or median values for each of the cues when they were available, in the most precise manner possible (e.g. separating mothers and fathers when possible) in all studies that contained IDS and ADS samples from the same talkers, as these constitute paired observations. Third, we collected results of any inferential statistical analyses that were carried out taking speakers as the units of analysis. When summarizing our results, we focus on main effects of register (to report any enhancement present in IDS compared to ADS), and their interaction with position when available (to test the increased discriminability of boundaries in IDS).

A. Results

1. Qualitative information

Our review revealed several important aspects of the methodology used in previous work. First, in this literature, a number of different criteria have been used to determine whether a boundary is present or not. The most common one was the presence of a pause that was at least 0.3 s in length (Stern *et al.*, 1983; Grieser and Kuhl, 1988; Fernald and Simon, 1984; Fernald *et al.*, 1989; Phillips, 1994; Church, 2002). Other papers decided that a boundary was present by relying on the syntax or meaning either exclusively, or primarily (i.e. relying on prosody when in doubt) (Kondaurova and Bergeson, 2011).

Second, the studies reviewed tended to have small sample sizes, with most studies representing each speaker and register through measures gathered from 100 tokens (e.g. number of breaks where pause was measured) or about 2 minutes of recordings; and the median number of speakers was 15.

Finally, among the studies we included, all but four focused exclusively on American English. The exceptions are one study on British English (Stern *et al.*, 1983), one involving German (Fernald and Simon, 1984), one study

TABLE I. Defining characteristics of the 36 paired IDS-ADS samples. Legend indicates the code that is used in Figure 1. Source codes the reference to the papers included in the review. Characteristics provides the total number of participants (collapsing across samples within a paper), additional details common to all samples, and discriminating characteristics between samples within a given paper. AE stands for American English; AAE African American English; BrE British English.

Legend	Source	Characteristics
B06-1 to 3	Bergeson *et al.* (2006)	27 AE mothers; samples differ in age (3-37 months) and hearing status
F84	Fernald and Simon (1984)	24 German mothers of newborns
F89-1 to 12	Fernald *et al.* (1989)	60 parents of 6-month-olds; samples differ in speaker sex and language
G88	Grieser and Kuhl (1988)	8 Mandarin Chinese mothers of 2-month-olds
K11-1 to 5	Kondaurova and Bergeson (2011)	27 AE mothers; samples differ in age (5-30 months) and hearing status
K12	Kondaurova *et al.* (2012)	28 AE mothers of infants 6-22 months
P94-1 to 2	Phillips (1994)	40 mothers of 6-8mo; samples differ in language (AE and AAE)
P10-1 to 2	Payne *et al.* (2010)	6 mothers of 2-year-olds; samples differ in language
R86-1 to 3	Bernstein Ratner (1986)	9 AE mothers; samples differ in infant production stage (e.g. first words)
S83-1 to 4	Stern *et al.* (1983)	6 BrE mothers; samples differ because IDS gathered longitudinally at birth, 4, 12, and 24 months (ADS gathered only once)
W15-1 to 2	Wang *et al.* (2014)	20 AE mothers; samples differ in age (4, 11 months)

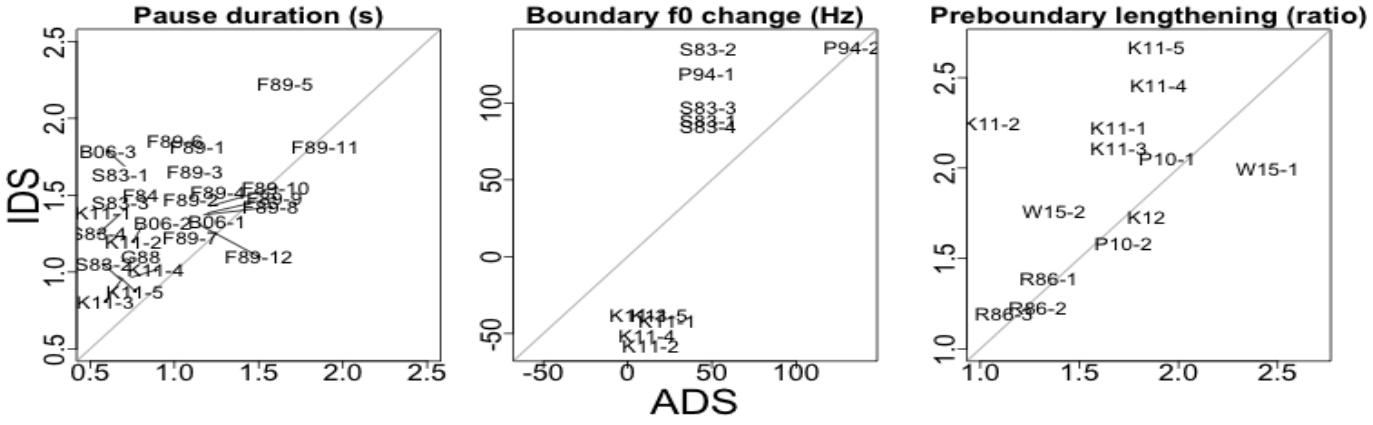

FIG. 1. Paired average values (IDS/ADS) for each of the main three acoustic cues of prosodic boundaries reported in previous work (see Table I for details on the samples). The plotted diagonal represents the equal value line between IDS and ADS.

on Mandarin (Grieser and Kuhl, 1988), and another one with 6 typologically-diverse linguistic populations (Fernald *et al.*, 1989). The latter would be very rich for the goal of documenting universal features, but unfortunately it reported information only on pauses (and this only at boundary locations).

2. Quantitative information: Paired IDS-ADS measurements

Taking all cues together, we were able to retrieve and combine paired IDS-ADS values from 9 papers, collectively reporting on a total of 36 samples (i.e., participant groups, typically having a homogeneous age and language). The defining characteristics of the 36 samples are provided in Table I, while differentiating characteristics of each sample are given in Table S3 of the Supplementary Material. Some of the samples contributed data for more than one cue, as noted below. Overall results are represented in Figure 1, which shows values in IDS as a function of ADS for each cue and paired observation.

It is apparent from the leftmost panel in Figure 1 that pauses at known boundary locations are longer in IDS than ADS. Indeed, despite a diversity of tasks, ages, criteria, etc., all paired data points are well above the diagonal line which indicates equal durations in IDS and ADS. For pitch, the results are less clear. Although we selected studies that used similar methods, paired values within each paper cluster together, and there are marked differences between studies with some of them finding larger values for IDS and others larger values for ADS (see middle panel in Figure 1). Finally, as with the results for pause duration, most of the results are above the diagonal for nucleus duration, and exceptions are at or just below it, consistent with larger correlates for IDS than ADS.

3. Quantitative information: Inferential analyses

The inferential statistics that we could retrieve from the studies are summarized in Table II. Four conclusions can be drawn from them. First, the mean values of the three cues are equally likely to be significantly affected

TABLE II. Results of inferential analyses reported in previous corpora studies (sorted by year of publication), for each of the three cues of prosodic boundaries. A field contains only R if a study tested for a main effect of register (typically only at boundary locations), and both R and P if both register and position (boundary, non-boundary) were taken into account. Factors and interaction between parentheses were found not to be significant, while those indicated with a plus sign were marginally significant. In Phillips (1994), AE stands for American English and AAE for African American English.

Reference	Test	Pause	Pitch	Nucleus
Stern *et al.* (1983)	ANOVA,Wilcoxon	(R)		
Fernald and Simon (1984)	Wilcoxon	R		
Bernstein Ratner (1986)	t-test			RxP
Fernald *et al.* (1989)	MANOVA,t-test	R		
Grieser and Kuhl (1988)	ANOVA,t-test	R^+		
Phillips (1994) AE	ANOVA		R	
Phillips (1994) AAE	ANOVA		(R)	
Church (2002)	ANOVA			R
Bergeson *et al.* (2006)	ANOVA,t-test	R		
Kondaurova and Bergeson (2011)	ANOVA	R	R, P, RxP	R, P, RxP
Kondaurova *et al.* (2012)	ANOVA			R, P, (RxP)
McMurray *et al.* (2013)	Mixed model			R, P, (RxP)
Wang *et al.* (2014)	Mixed model		(R), (P), (RxP)	(R), P, (RxP)

by register. Indeed, in 11 out of the 15 cases in which the main effect of register was probed, a significant or marginally significant effect of register was found. Second, there is very little work directly assessing whether the presence of a prosodic break (i.e. position) interacts with register: a total of 7 interaction terms were found in our review. Third, current evidence regarding the possibility of an interaction is far from clear: only 3 of those 7 tests were significant. Finally, only one study (Kondaurova and Bergeson, 2011) has provided a comprehensive account, inspecting all three types of cues together, and comparing them across registers.

B. Discussion

The systematic review revealed substantial heterogeneity in the studies comparing IDS and ADS. However, a general trend emerges: of the three cues, pause duration is almost always systematically larger in IDS than ADS. Nucleus duration is also generally found to be larger in IDS than ADS. However, for f0 change, the results are inconclusive, with half the studies (Stern *et al.*, 1983; Phillips, 1994) finding a larger degree of reset in IDS, and the other half an effect in the opposite direction (Kondaurova and Bergeson, 2011). Since the three cues do not necessarily behave in the same way across studies, this raises the issue of whether they are consistent enough for an infant to be able to use them in order to learn prosodic boundaries, and whether the resulting algorithm would still favor IDS over ADS.

In addition, most previous studies measured the value of the cue at the location of a prosodic boundary, but rarely measured the same cues in non-boundary positions. In order to address the issue of the learnability of prosodic boundaries, however, one must analyze and report information in both boundary and non-boundary cases. Unfortunately, only a handful of corpora analyses have reported on the contrast between boundary and non-boundary features, and those that do mostly focus on a single cue. Among this handful, only half of the interaction terms were reported to be significant (Bernstein Ratner, 1986; Kondaurova and Bergeson, 2011). Thus, the hyperspeech hypothesis is only very partially addressed by published research, and with mixed results.

Furthermore, we estimate the average recording length in these studies to be about 1 hour (collapsing across registers and participants); often, only a subset of data with a duration between 2 and 4 minutes per participant has been inspected. It is unclear how large a recording ought to be in order to accurately represent the talker and register. We know of no data that would enable one to compute the minimum sample size for stable results in this domain, but given the variability across studies that we documented, a larger data set should, if anything, clarify the picture.

Finally, our systematic review revealed an important problem regarding the definition of prosodic breaks. Indeed, much previous work required the presence of a pause to establish that a break was present, which does not address the question of whether pause is a necessary cue for a break. Other studies employed syntactic criteria, which avoids this problem but introduces uncertainty regarding whether a break was truly present, since syntactic breaks are not necessarily accompanied by prosodic breaks. In the following work, we use the Tones and Break Indices (ToBI, Silverman *et al.*, 1992) framework for prosodic annotation, which is an attempt at standardizing the marking of prosodic breaks across several languages, using experts' judgments based on a range of phonological and syntactic criteria.

III. CORPUS ANALYSIS

Given that previous results are mixed and scarce, we decided to carry out our own corpus analysis, on a large dataset of Japanese infant- and adult-directed speech. We aim to provide a descriptive account of the hyperspeech hypothesis for prosodic boundaries. This involves

both establishing the main values and the variability of the prosodic boundary cues across participants in ADS and IDS, and determining the systematicity with which these cues vary across registers. Our corpus analysis improves upon previous studies in four salient ways.

First, we used break indices from the X-JToBI framework (a version of ToBI adapted for use with Japanese; Maekawa *et al.*, 2002). In this framework, expert coders make a decision regarding the presence of disjuncture between every pair of words; if there is a perceived break, they classify it in terms of its perceived importance. In this process, coders use all sources of phonetic information, including a number of language-specific cues, to decide the strength of a break, from high-level prosodic patterns (e.g. the fact that certain tones only occur at certain prosodic positions) all the way down to allophonic patterns (e.g. in English, the break index between the words 'did you' might be 0 if there is palatalization of the "d", and 1 otherwise). The X-JToBI labeling standard defines four main break levels (labeled between 0, the weakest, and 3, the strongest level), along with some intermediate levels between these four and special symbols marking disfluency phenomena. For the present work, we considered as prosodic boundaries all X-JToBI break levels 2 (accentual phrase boundary) and 3 (intonation phrase boundary), as well as locations where the coder was uncertain between these two levels, marked in our corpus with the label 2+. These boundaries roughly correspond to the phonological and intonational phrases of the prosodic hierarchy. By employing this definition of prosodic boundaries we can perform a more precise analysis of the acoustic cues involved, because under our definition, boundaries are not limited to those marked by a pause [3] or aligned with a syntactic break.

Second, we considered *all* syllable sequences. As mentioned briefly above, other studies have compared phrase-final syllables to phrase-medial ones (Bernstein Ratner, 1986) or phrase-final syllables to the onset syllable of the next phrase (Kondaurova and Bergeson, 2011). The latter method is somewhat closer to how a perceiver would face the task of boundary detection on the fly, with the limitation that only syllables at a boundary can be studied. We generalized this method by considering a running window of two syllables throughout all phrases. When the pair of syllables spanned a boundary as defined above, then it was marked as a boundary; otherwise it was classified as a non-boundary case.

Third, we used a larger corpus than previous studies. Our corpus included about 40 minutes from each of 22 mothers, totalling 14 hours of speech. This is more than a 10-fold increase in number of measurements compared to the past studies reported in our review (e.g. Bergeson *et al.*, 2006 and Kondaurova and Bergeson, 2011 analyzed 54 minutes each). Including 22 participants also allowed us to measure individual variation in ADS versus IDS.

Finally, we investigated all three main cues to prosodic boundaries in both boundary and non-boundary position, and assessed their potential interaction with register (IDS versus ADS). As mentioned in the systematic review, only one previous study has done this, and it focused on American English. Our study thus represents an extension to a language with a different prosodic organization, as explained in the next subsection.

A. Japanese prosodic organization

In Japanese, speech is grouped into two levels of prosodic phrasing, called accentual and intonation phrases, defined both tonally and on the basis of the degree of perceived disjuncture (Venditti, 2005). Additionally, phrases can be marked by optional boundary pitch movements, which convey pragmatic meaning (e.g. questioning).

Accentual phrases are tonally defined as sequences of words 1) having a pitch rise on the first or second mora of the sequence, with the pitch then falling towards the end of the accentual phrase, and 2) containing at most one lexical pitch accent, the location of which determines precisely where the pitch fall occurs. Pairs of words within the same accentual phrase will have a lower perceived disjuncture than pairs of words spanning an accentual phrase boundary. At the level of the accentual phrase we encounter the phonological process of "downstep", by which the pitch height of each accentual phrase decreases if it was preceded by another lexically accented phrase. The beginning of an intonation phrase is characterized by the speaker setting a new pitch range, independent from the pitch specification of the previous intonation phrase. In terms of the perceived disjuncture, intonation phrases are defined as having a stronger disjuncture between words spanning their boundaries than for words within or across accentual phrases.

Similar to other languages, Japanese prosodic boundaries are marked acoustically by the three cues investigated in our literature review: pause duration, nucleus duration and f0 reset (see Vaissière (1983) for a review). The Japanese ToBI system uses these cues, along with others like f0 lowering or voice quality, to make judgments on the perceived disjuncture between adjacent words (Venditti, 2005). As mentioned above, the Japanese prosodic hierarchy can also be described in terms of tonal movements (Beckman and Pierrehumbert, 1986) (e.g. an L boundary tone marking the end of an accentual phrase). The role of these three features in marking prosodic boundaries is not limited to ADS: comparing acoustic cues at boundary and non-boundary position in infant-directed speech, Fisher and Tokura (1996) have shown that pauses, final lengthening and pitch excursion are larger just before a prosodic boundary than in a non-final position. Furthermore, these cues were also shown to be useful for the automatic detection of prosodic boundaries: f0 contours (Campbell, 1996), pause duration along linguistic information (Akita *et al.*, 2006) or a combination of the three acoustic features investigated here (Ludusan *et al.*, 2015).

B. Materials

We used the RIKEN corpus, which contains 14.5 hours of both infant-directed and adult-directed speech

TABLE III. Summary of the statistical analysis performed on the corpus, for the three cues reported in the literature: pause duration, nucleus duration and f0 change. For each speech register (ADS, IDS) we report the mean and standard deviation of each cue, the boundary/non-boundary t-test value, and the t-test results for the interaction between the two registers ($^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$; $df = 21$ for each analysis).

Feature	ADS						IDS						Interac. t-test
	Boundary		Non-boundary		Diff		Boundary		Non-boundary		Diff		
	Mean	Stdev	Mean	Stdev	Mean	t-test	Mean	Stdev	Mean	Stdev	Mean	t-test	
pause duration	0.922	0.591	1.012	0.671	-0.090	-1.46	1.009	0.623	0.881	0.571	0.128	8.40^{***}	-5.11^{***}
nucleus duration	0.136	0.129	0.087	0.094	0.049	14.63^{***}	0.165	0.203	0.097	0.102	0.068	10.57^{***}	-4.11^{***}
f0 change	7.617	94.71	-2.478	89.03	10.10	4.66^{***}	1.200	129.2	-0.474	102.4	1.674	0.97	2.57^{*}

(Mazuka *et al.*, 2006). Approximately 11 hours consists of interactions between 22 Japanese mothers and their 18- to 24-month-olds, while the remainder consists of conversations between the same mothers and an experimenter. The corpus was fully transcribed and hand-annotated with segmental labels, morphological information, and intonation labels. For the analysis performed here we have considered the entire vowel (including the long mora, coded as H) as the syllabic nucleus. We treated the moraic nasal (coded as N) as a syllable coda, excluding it from any measurements pertaining to the syllable nucleus.

The corpus was annotated for prosody using the X-JToBI standard (Maekawa *et al.*, 2002). The prosodic annotation was done mostly by one well-trained X-JToBI labeler, who was assisted by two other experts. In order to ensure a higher annotation consistency, the first expert double-checked the portion of the corpus coded by the other two experts. The coders marked the boundaries according to the X-JToBI guidelines: prosodic phrases where the pitch range is reset were marked as intonation phrases, while phrases where an initial pitch rise occurs were considered to be accentual phrases (Mazuka *et al.*, 2006). Intonation phrases contain one or more accentual phrases, with the latter containing one or more words. While no inter-annotator reliability was computed for the RIKEN corpus, the X-JToBI prosodic phrasing annotation has been reported to result in relatively high reliability (Cohen's Kappa $\kappa = 0.73$, on a subset of the Corpus of Spontaneous Japanese, as reported in Igarashi *et al.*, 2013).

In this study, we focused on breaks coded either as level 2 or level 3 or as an intermediate level between 2 and 3. This resulted in a total of 9,245 boundaries for the ADS register and 22,303 boundaries for the IDS register.

The following cues were used for the corpus analysis: *pause duration*, defined as the duration of any silent pause following the current syllable (seconds); *nucleus duration*, being the duration of the current syllable nucleus (seconds); and *f0 change*, representing the difference between the average f0 value of the following syllable and the average f0 of the current syllable (hertz). These cues were chosen because they have already been used in the literature on IDS: pause has been investigated in many previous studies (Stern *et al.*, 1983; Fernald *et al.*, 1989; Bergeson *et al.*, 2006); as has nucleus duration (Bernstein Ratner, 1986; Kondaurova and Bergeson, 2011; Kondaurova *et al.*, 2012) and f0 change (Stern *et al.*, 1983; Phillips, 1994; Kondaurova and Bergeson, 2011).

For the pause duration cue, we excluded from our analysis all inter-speaker pauses, which we define as periods during which the mother is not speaking but the infant or the experimenter is. While the IDS sub-part of the corpus is marked with the infant's vocalizations, making this task easy, the ADS sub-part was not coded for the times when the experimenter was speaking. In order to determine which pauses contained non-mother speech in the ADS data, we used the voice activity detector (VAD) included in Praat (Boersma and Weenink, 2014) to obtain an approximation of the experimenter's speech intervals. Several values for the VAD threshold were tested and the best value obtained was used to run the VAD on the entire ADS corpus. The final timings for the experimenter's speech were determined on the basis of a visual inspection of the results obtained from the VAD against the speech waveform, performed by one of the authors. In this way we were able to exclude both IDS pauses which contained infant speech and ADS pauses which contained experimenter speech, as determined by the Praat VAD. We also excluded from the pause duration analysis all the syllables followed by a pause shorter or equal to 300 ms (Stern *et al.*, 1983; Grieser and Kuhl, 1988; Fernald *et al.*, 1989) as well as by pauses longer than 3 s (Stern *et al.*, 1983; Grieser and Kuhl, 1988; Kondaurova and Bergeson, 2011). We chose to impose a lower limit on the pause duration due to the presence of very long geminate stop closures in Japanese which might be confounded with pauses (Kawahara (2015) reports closures longer than 200 ms, on average, in several studies). The upper limit was set in order not to consider intervals of time in which the mother was pausing or was interacting non-verbally with the infant. As a result, out of a total of 4,262 boundary pauses in ADS, 407 were removed due to overlap with speech and 1,447 excluded for being shorter than 300 ms or longer than 3 s, leaving 2,408 boundary pauses to be analyzed. The numbers (total - excluded for overlap - excluded for length = analyzed) were 1,293-112-616=565 for non-boundary pauses in ADS, 14,234-3,239-3,585=7,410 for boundary pauses in IDS and 2,618-387-1,105=1,126 for non-boundary pauses in IDS, respectively.

C. Methods

We performed a statistical analysis by first averaging the value of each cue across all syllables for a given speaker, boundary condition and register separately. Second, for each feature and register, we applied paired t-tests across speakers comparing the boundary and non-boundary conditions. Finally, for each feature, the interaction between boundary condition and register was computed as a t-test on the ADS minus IDS mean feature value.

D. Results

The results of this analysis are illustrated in Table III. We can observe that all features, except pause duration in ADS and f0 change in IDS, are significant for the boundary/non-boundary distinction. In addition, the interaction between the two speech registers was found to be significant for all three cues, although in different directions: for pause and nucleus duration, the effect of boundary was larger for IDS than for ADS, while the converse was true for f0 change. These effects were modestly to strongly consistent across speakers depending on the cue. Indeed, a greater distinction in IDS than ADS for pause duration was observed for 21 out of 22 speakers; for nucleus duration in 19 out of 22 (with two more showing the opposite trend, and the last one no difference between IDS and ADS); but for f0 change, in 15 out of 22 speakers the distinction was greater in ADS than IDS (with the remaining 7 showing the opposite trend). These results are presented in more detail in Figure S1 of the Supplementary Material.

E. Discussion

In an analysis of a corpus of spoken Japanese, we found that each of the three cues to prosodic breaks (pause duration, nucleus duration, and f0 change) were numerically larger across boundaries than within prosodic units (with the sole exception of pause duration in ADS), which suggests that a learner could use such cues to detect boundaries. However, one cue (nucleus duration) was consistent across speakers, while others (pause duration and f0) were weaker or more variable. We also found that for the first two cues, the difference between boundary and non-boundary was larger in IDS than in ADS, suggesting that IDS might be easier to segment than ADS, in line with the hyperspeech hypothesis.

Surprisingly, the pause results were significant only for IDS. The lack of significance in ADS was due to the fact that the this register contained filled pauses and disfluencies (marked in our corpus with the labels F, D, and PB thus a non-boundary) followed by longer pauses[4], which resulted in pauses at non-boundary being longer than pauses at boundary in ADS[5].

Because register interactions in the f0 cue went in the opposite direction to those found for the nucleus and pause cues, it is conceivable that combining all three cues in a single boundary-finding mechanism would result in these differences cancelling each other out to a certain extent. Also, while we showed the reliability of the nucleus cue across speakers, it is important to also test its reliability across trials (syllable samples). Indeed, the prosodic boundary detection algorithm must make a correct decision for any given sequence of two syllables, not on the basis of average values across sentences. We will address these two issues in the next section, using machine learning techniques.

IV. MACHINE LEARNING EXPERIMENTS

The review and corpus analysis reported above revealed that the three cues traditionally studied in relation to prosodic boundaries in IDS and ADS (pause duration, nucleus duration and f0 change) are not necessarily of the same strength and reliability, and that they interact with register to different extents and sometimes in unexpected ways. Therefore, in order to test the hyperspeech hypothesis for prosodic constituents, it is necessary to use a more integrated approach in which the cues are combined in order to build a prosodic boundary classifier.

We use machine learning tools to address this issue in two ways. First, we use two supervised learning methods (decision trees and neural networks) in order to determine the optimal cue combination for the purpose of prosodic boundary classification. We train a classifier to distinguish boundaries from non-boundaries using the X-JToBI labels on a subset of the data, and then test the resulting classifier on unseen data (using a 10-fold cross-validation scheme). We derive a different classifier for each register, in order to determine the maximal performance attainable with this type of dataset. The contribution of the different acoustic cues is then analyzed separately in order to establish cue weighting for each register. Second, we run an unsupervised learning algorithm on the same data, again deriving a separate classifier for each of the two registers. This unsupervised method directly addresses the hyperspeech hypothesis by simulating the infant, who has access only to the acoustic data and no knowledge of the boundary labels.

A. Methods

1. Materials

The same dataset was used as in the corpus analysis. For the learning experiments, the IDS and ADS recordings from two speakers were used as a development set, while the rest of the recordings (from 20 speakers) were used as the evaluation set. The development set contained 2,373 syllables for ADS and 6,154 syllables for IDS, while the evaluation set had 35,983 and 73,730 syllables for ADS and IDS, respectively. The number of prosodic boundaries present in the evaluation set was 8,633 for ADS and 20,506 for IDS.

2. Acoustic features

The same acoustic features were employed as in the previous sections: *pause*, *nucleus duration* and *f0 change*. They have previously been employed successfully in the boundary detection literature, together with other features (Ludusan and Dupoux, 2014).

For the learning experiments, the *pause* feature is slightly different than the pause duration employed in the statistical analysis, representing a combination of pause duration and a categorical definition of pause, defined only by the presence or absence of pause (see section S2 of the Supplementary Material for a statistical analysis of this categorical pause definition). Pauses were normalized between 0 and 1, with the 0 level corresponding to a pause of 300 ms or shorter, while 1 represents a pause of 3 s or longer. Intermediate pause values were obtained through a linear transformation, while pauses shorter than 300 ms were set to 0, and pauses longer than 3 s were given the same value as a 3 s long pause (equal to 1). This normalization allows us to avoid both missing values and the distribution with a long tail typical of pauses.

3. Learning algorithms

As learning algorithms to be applied to the extracted features, we chose two supervised methods—decision trees and neural networks—and one unsupervised method based on Gaussian mixtures with the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). Decision trees were chosen because their structure can give insights into the usefulness of each individual cue as well as their interactions, while neural networks have previously been used successfully for prosodic boundary detection (Ananthakrishnan and Narayanan, 2008).

The implementations of the algorithms used in this paper are part of the Weka toolbox (Hall *et al.*, 2009). The EM clustering algorithm models the input attributes by means of a mixture of Gaussians with a diagonal covariance matrix. The number of clusters to be found by the model was fixed at 2, corresponding to the boundary/non-boundary cases. In the training phase, the EM clustering was run for a maximum of 100 iterations, using the improvement in log likelihood as a stop criterion. The minimum log likelihood improvement required to perform another iteration of the E and M steps was set to 1E-6. Once the model was built, during the test phase the clustering algorithm returned class probability values for each instance, which were then used for computing the evaluation metric.

The decision trees implement the C4.5 algorithm (Quinlan, 1993), with pruning. During training, the trees are generated incrementally from the root node towards the leaf nodes, by finding, at each node of the tree, the attribute having the highest information gain ratio. Once this attribute is determined, the set of samples at the current node are split into two child nodes, according to the attribute value tested. The process is repeated for the obtained nodes until the resulting child nodes belong to the same class and a leaf node is created with that given class label. Thus, one label can appear at multiple leaf nodes, depending on the attribute tested in each node. At evaluation time, samples from the test set are checked against the decision criteria in each node and are given the label corresponding to the leaf node reached. We decided to use pruned trees because the pruning process decreases the risk of overfitting the model on the training data. Pruning is carried out by replacing unreliable non-terminal nodes with leaf nodes, based on statistical confidence estimates. The main parameters of the implementation used, the pruning confidence factor, and the minimum number of instances in a leaf node were obtained on a development set by maximizing the classification performance. The pruning confidence was varied between 0 and 0.5, with lower values associated with more pruning. The minimum number of instances in a leaf node also plays a role in the complexity of the tree and was varied between 1 and 50.

The neural network used was a three-layered feedforward perceptron (Rumelhart *et al.*, 1985) implemented in Weka. It had an input layer with a number of nodes equal to the number of features employed (three), an output layer with two nodes (corresponding to the number of classes) and one hidden layer. The nodes in each layer use a sigmoid activation function and are fully connected with those in the following layer. The number of nodes in the hidden layer was obtained by optimizing the learning performance on the development set, and varied between 0 and 50. A higher number of nodes translates into a more complex network, with a longer training time. The cost function, minimized through stochastic gradient descent using the backpropagation algorithm, was the squared error. The learning rate and the momentum rate of the backpropagation algorithm were both set to 0.1. The training was stopped when 20,000 epochs were reached or when the error on a subset (20%) of the training material rose 20 times in a row.

B. Procedure

We applied the three learning algorithms to the feature set described above. The best parameters for each supervised algorithm were determined on the development set, while the results reported in this section were obtained on the evaluation set. The same experimental setting was used for all learning algorithms: a per-speaker evaluation was performed and the average of the individual results calculated. Only the procedure that was applied differed between types of algorithms: the supervised methods used 10-fold cross-validation, while the unsupervised method was run directly on the evaluation set.

The obtained classes/clusters were evaluated against the reference boundary labels by computing the area under the receiver operating characteristic (ROC) curve (AUC). The ROC curve is obtained by varying the threshold for the class probability estimates and plotting for each value the resulting true positive rate versus the

false positive rate. The AUC can be interpreted as the probability of making a correct choice in a forced choice task where one is given a random pair of tokens, one instantiating a boundary, and the other a non-boundary. The chance level for the AUC is therefore 0.5. Using AUC as an evaluation measure is especially useful when comparing two databases of different sizes and distributions of boundaries and non-boundaries, as in the ADS and IDS sub-parts of our corpus.

The following parameter values were obtained on the development set for decision trees (pruning confidence/minimum number of instances in a leaf node): (0.25/6) for IDS and (0.4/6) for ADS. The optimum number of nodes in the hidden layer of the neural networks was determined to be 9 for IDS and 28 for ADS.

C. Results

1. Supervised and unsupervised learning

Figure 2 shows a comparison of the results obtained for the two registers and the different machine learning algorithms employed. The bottom line represents the chance level, equal to 0.5 AUC. The error bars correspond to the standard deviation across the speakers in the evaluation set. It appears that IDS boundaries are consistently better learned than ADS boundaries, for all features and all learning algorithms. The improvement ranges between 2.3% (EM) and 7.7% (decision trees), with an average gain across algorithms of 5.8%.

A more detailed analysis of the results shows that the advantage of IDS over ADS, for both unsupervised and supervised learning, can be seen not only in the average performance across the speakers, but consistently across individual speakers. For each of the three learning algorithms we obtained a higher performance in IDS than ADS for all but one speaker (for more details, see the Supplementary Material, Figure S2).

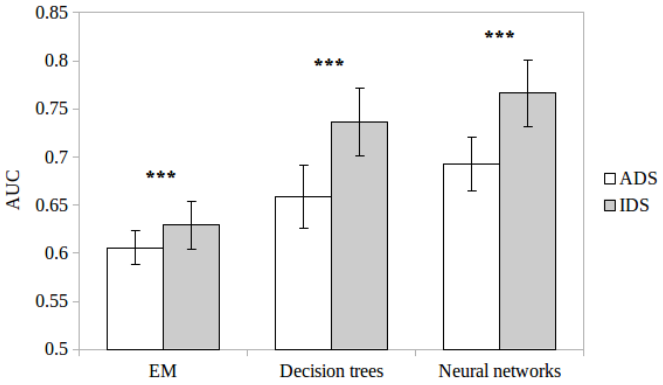

FIG. 2. Area Under the ROC Curve results obtained for the classification of boundary versus non-boundary syllables on the two speech registers (ADS and IDS), using both unsupervised (EM) and supervised (decision trees and neural networks) learning algorithms. The error bars represent the standard deviation across the 20 speakers in the corpus. A t-test showed a highly significant difference between the two registers, for all learning algorithms ($^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$).

2. Analyzing cue importance

Our goal is to analyze the relative importance of the three acoustic cues, as determined by their role in the learning process. To do this, we first examined the learned decision trees and extracted the maximum level of the tree in which each feature appeared, with level 1 representing the root node. Using the level of the decision tree as a proxy for the importance of each cue seems natural given the tree's hierarchical structure. The results are presented in Table IV, together with a denominator corresponding to the depth (maximum number of levels) of the obtained tree. It revealed some interesting discrepancies with respect to the results of the statistical analysis. For instance, *pause duration* in ADS and *f0 change* in IDS were not found to differ significantly in boundary versus non-boundary positions, yet they appeared in the decision tree nonetheless. Our pause feature appeared for both registers in the root node, outranking the other two cues. As for *f0 change*, an analysis showed that its value goes in opposite directions for accentual and intonation phrase boundaries, respectively, with the overall average approaching zero (see Table S5 in the Supplementary Material). But when used in a decision tree, this cue can still help discriminate when used with different decision values.

TABLE IV. Summary of the cue importance analysis performed. For the two speech registers (ADS, IDS) we illustrate the maximum level achieved by each of the three cues in the decision tree (e.g. 2/6 means that the cue first appeared on the second level of a tree having a depth of six). The depth of the tree can be higher than the number of features as cues may be reused on lower levels of the tree, with different values tested in the decision.

Feature	ADS	IDS
pause	1/6	1/4
nucleus duration	2/6	2/4
f0 change	3/6	3/4

While statistical analyses can be used to describe the cue patterns found in the data, learning algorithms are needed to tell us how these cues can be optimally integrated. In order to respond to this question, we investigated how our best learning system (neural networks) performed when presented only with a subset of the three acoustic cues. We gave in input to the learning algorithm single cues as well as all combinations of two cues, and compared the results with those obtained when all three cues were used. Models were trained and evaluated for each condition using the same procedure as in IV.B. Table V illustrates the results for all combinations of the two registers. All significant differences are due to better performances with IDS than ADS.[6] Interestingly, all the combinations which show improvements in IDS over

ADS include pause. This suggests that the main cue for boundary marking in IDS is the pause.

To further investigate the role of each cue, we performed pairwise significance tests between the results of each feature/combination of features presented in Table V, for each register separately (see Table S6 in Supplementary Material). It revealed that adding an extra cue to any of the cue combinations (e.g. adding f0 change to pause, or adding pause to a combination of nucleus duration and f0 change) improved performance significantly in IDS in all cases but one (f0 change added to a combination of pause and nucleus duration had a marginal effect, $p = 0.075$). An important conclusion can be drawn from this: the three cues examined in this study represent mutually complementary sources of information and are all useful for learning, regardless of how the descriptive statistics have characterized them.

TABLE V. Area Under the ROC Curve results obtained for the classification of boundary versus non-boundary syllables on the two speech registers using the best algorithm (neural networks), and testing different combinations of cues (P = pause; N = nucleus duration; F = f0 change). A t-test determined the significance of the difference between registers ($^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$).

Feature	ADS	IDS	Register (t-test)
P	.643	.741	-12.40 ***
N	.632	.616	1.93
F	.552	.539	1.07
PN	.693	.763	-9.23 ***
PF	.658	.751	-9.13 ***
NF	.645	.647	-0.16
PNF	.693	.766	-8.06 ***

D. Discussion

The results of the machine learning study show that irrespective of the choice of learning algorithms, there are significant learnability differences, as large as 8% AUC, between IDS and ADS. These differences were found to be highly significant for each of the learning algorithms tested. Furthermore, we have seen that prosodic boundaries are better learned in IDS than ADS (see section S3.2 of the Supplementary Material for a demonstration that this statement holds even when less data is given to the learning algorithm). These findings confirm the hyperspeech hypothesis for prosodic boundaries.

In addition, our results underline the importance of a learning-based approach, which must complement any statistical analyses performed on acoustic cues. Indeed, we observed several salient discrepancies between the results of the statistical analysis and the use of the features in the decision trees. Features that were non-significant in the statistical analysis were still found to be relevant in the decision trees (*pause* in ADS and *f0 change* in IDS). This shows that, in a learning environment, the interaction between different cues may change the importance of individual cues.

V. GENERAL DISCUSSION AND CONCLUSIONS

This study examined the long-standing claim that prosodic boundaries are more easily detected in IDS than ADS. Our systematic review revealed that, among the acoustic cues that have been most studied, pause duration is consistently larger in IDS compared to ADS, as is nucleus duration, whereas the results on f0 change vary across studies. We also found that earlier studies rarely included non-boundaries in their analysis, making it difficult to know whether the larger values found in IDS were not simply a side effect of slower speech rate, and thus not necessarily enhanced as cues to the presence of prosodic boundaries. Our own corpus analysis on a larger dataset than previously studied revealed that pause duration and nucleus duration were reliable boundary cues and they were enhanced in IDS compared to ADS, while f0 change was less reliable. Previous work on Japanese suggests that prosodic boundaries are marked by the three acoustic cues, both in ADS (Vaissière, 1983) and IDS (Fisher and Tokura, 1996), with pause duration being a stronger boundary cue in IDS than ADS (Fernald *et al.*, 1989). Our study, the first in Japanese to look at all three cues in both registers, confirms this statement, except for pause duration in ADS and f0 change in IDS not being significant. We can moreover add to previous work that all three cues show interactions with register, two that are in line with the hyperspeech view of IDS and f0 change patterning in the opposite direction. Finally, we ran machine learning algorithms, both supervised and unsupervised, on the same corpus using three algorithms, and found that, in all instances, IDS boundaries were easier to learn than ADS boundaries.

The claim that parents may facilitate infants' language learning through the characteristics of the IDS register (the hyperspeech hypothesis) has been found to be controversial at a number of phonological levels (see Cristia, 2013, for a review). We have demonstrated that, at least in the domain of prosodic boundaries, this hypothesis can be strongly supported when the appropriate analysis is carried out. Our corpus of child-directed Japanese speech is particularly interesting because acoustic measurements taken from it have refuted predictions from the hyperspeech hypothesis for sound categories (Martin *et al.*, 2015). In other words, the same parents who make sound categories more difficult to discriminate in IDS than ADS (presumably because of increased variability), make their prosodic boundaries easier to discriminate.[7] We would like to point out, however, that our study does not demonstrate that the prosodic enhancement present in IDS is actually used by infants (i.e. we do not demonstrate that parents' facilitation is effective). In order to do this, we would have to study individual variability and show, for example, that the level of prosodic break discriminability present in the adults' speech correlates with their infants' learning.

Regarding the cues that are important for learning prosodic breaks, it is noteworthy that, in our study, the acoustic cue which is consistently, across all conditions and analyses, ranked the most important is the *pause*. Furthermore, if we were to consider every time the

speaker stopped talking for more than 300 ms as a pause, and adopt a very simple system which placed a prosodic boundary after every syllable succeeded by a pause, we would obtain an AUC of 64.6% for ADS and 73.6% for IDS, which is well above chance (50%). The use of supervised algorithms and additional features (*nucleus duration* and *f0 change*) increased performances only slightly, to 69.3% and 76.6% for ADS and IDS, respectively. Of course, it is important to note that the experts who coded the breaks may have been using pause as a criterion, so we cannot completely escape the circularity noted in section II. One solution to this problem might be to stop using human labels as the gold standard, and instead use the effect of boundaries on the learning of some other linguistic level(s) (e.g. word segmentation as in Ludusan *et al.* (2014, 2015)).

It is worth mentioning, however, that pause is not the only factor at play. Even f0 cues, which were also shown to be only weakly effective in the statistical analysis, play a role in the learnability results, presumably through non-linear effects and interactions with other cues. This highlights the importance of using more complex statistical models or machine learning tools in order to assess the value of particular acoustic cues in the learning of linguistic structures. While we opted for the latter in this study in order to test the learnability of prosodic boundaries, the former option is also viable (see Werker *et al.* (2007) for the use of hierarchical multi-level logistic regression analysis for the learning of phonetic categories).

The detailed results regarding the importance of cues that we found here also need to be replicated in further studies using other languages. Recent laboratory work has began exploring cue weighting for boundary detection in infants learning different languages (so far, English, Dutch, and German; see a recent review in Wellmann *et al.*, 2012). Our machine learning results have revealed the optimal cue-weighting for this Japanese corpus, and thus cannot be compared directly to that work. Nonetheless, applying similar methods to other corpora may allow researchers to develop more precise predictions regarding which cues are valuable for the infant's linguistic environment, and thus confirm to what extent cue-weighting reflects optimal attunement to environmental input.

Cross-linguistic extensions would be particularly desirable given the variability in results found in the systematic review. Although to a certain extent this variation could be due to sampling errors associated with the use of small data sets and/or variation in the criteria used to define boundaries, it is entirely plausible that there are sizable cross-linguistic differences in IDS enhancement. Previous work at the segmental level suggests that this variation may not relate to obvious linguistic features of the language. Kuhl *et al.* (1997) were confident of the validity of the vowel triangle expansion they observed because they could replicate it in three typologically diverse languages, namely English, Russian, and Swedish. Yet Benders (2013), studying Dutch (a language that is historically and typologically similar to English), and Englund and Behne (2006), studying Norwegian (similar to Swedish), both observed a significant reduction of the vowel triangle in IDS compared to ADS.

Another dimension of variation that ought to be explored in future work is undoubtedly the addressee's age and/or level of development. We studied the speech of mothers whose normally-developing children were between 18 and 24 months of age. It would be interesting to know whether the same kind of enhancement exists both before and after that time window. Work on vowel triangle properties suggests a decrease in hyperarticulation between infancy and childhood (Liu *et al.*, 2009), and research on paralinguistic prosodic features suggests non-linear changes with age that are specific to each type of cue (Amano *et al.*, 2009; Stern *et al.*, 1983; Vosoughi and Roy, 2012). Furthermore, data in our systematic review reveals a non-significant linear trend for pause duration at the boundary to decrease across the age ranges studied (see Supplementary Material). It would be interesting to see whether (and how) prosodic boundary discriminability changes with infant age/developmental stages, and whether this time course relates to infants' learning profiles.

Analyses of different corpora containing speech directed to children of different ages, and who are learning different languages, are thus needed. We believe that such extensions will be most informative if they report not only statistical tests allowing the derivation of effect sizes (thus facilitating integration of results via a meta-analysis), but also results from supervised and/or unsupervised modeling approaches, which reveal the accuracy of a learner who takes into account multiple cues.

Acknowledgments

BL and ED's research was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), and AC by the Agence Nationale pour la Recherche (ANR-14-CE30-0003 MechELex). It was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the École des Neurosciences de Paris, and the Région Île-de-France (DIM cerveau et pensée).

1. Evidence that prosodic boundaries require learning comes from the study of the "closure positive shift" in electroencephalography, which has been shown to develop between 5 months and 6 years of age, with a gradual integration of acoustic cues other than pause (Steinhauer and Friederici, 2001).

2. We acknowledge that the importance of each of these cues for boundary perception has been found to vary across languages, tasks, and ages (see Schmitz (2008) for a comprehensive review). Since the evidence on this question is complex, and the present manuscript does not report on human perception, we do not delve into it extensively here, but merely indicate that any corpus study on prosodic boundaries must take into account these three acoustic cues.

3. Only 23.8% of IDS and 12.5% of ADS boundaries in our corpus are followed by a pause longer than 300 ms.

4. In fact, more than 97% of the non-boundary pauses correspond to instances of disfluency.

5. If one would relax the constraints imposed on the pause duration cue, by taking into account also pauses shorter than 300 ms, pause duration would be able to discriminate boundary from non-boundary cases also in ADS. Performing the same analysis on the unconstrained pause duration cue we obtained significant differences in ADS ($t = 2.55$, $p < 0.05$) and highly significant differences for IDS ($t = 15.52$, $p < 0.001$) and the interaction between registers ($t = -5.86$, $p < 0.001$).

6. However, it should be noted that one of the two cases showing an ADS advantage (when only nucleus duration is provided) is marginally significant, $p = 0.069$.

7. This result could indicate that, if IDS does in fact have a functional role, this role is geared towards helping infants acquire linguistic structures that depend on prosodic breaks (i.e. word boundaries or syntax), but not necessarily linguistic structures that depend on phonemes (phonetic inventory or lexical word forms).

Akita, Y., Saikou, M., Nanjo, H., and Kawahara, T. (**2006**). "Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines.", in *Proc. of INTERSPEECH*, 1033–1036.

Amano, S., Kondo, T., Kato, K., and Nakatani, T. (**2009**). "Development of Japanese infant speech database from longitudinal recordings", Speech Commun. **51**, 510–520.

Ananthakrishnan, S. and Narayanan, S. (**2008**). "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence", IEEE Trans. Audio Speech Lang. Process. **16**, 216–228.

Beckman, M. E. and Pierrehumbert, J. B. (**1986**). "Intonational structure in Japanese and English", Phonology **3**, 255–309.

Benders, T. (**2013**). "Mommy is only happy! Dutch mothers realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent", Infant Behav. Dev. **36**, 847–862.

Bergeson, T., Miller, R., and McCune, K. (**2006**). "Mothers' speech to hearing-impaired infants and children with cochlear implants", Infancy **10**, 221–240.

Bernstein Ratner, N. (**1986**). "Durational cues which mark clause boundaries in motherchild speech", J. Phon. **14**, 1303–1309.

Boersma, P. and Weenink, D. (**2014**). "Praat: doing phonetics by computer [Computer program]", `http://www.praat.org`, version 5.4.01, retrieved Nov-14.

Campbell, N. (**1996**). "Autolabelling Japanese ToBI", in *Proc. of ICSLP*, volume 4, 2399–2402.

Church, R. (**2002**). "Prosodic modifications in infant-directed speech", Ph.D. thesis, University of British Columbia.

Cristia, A. (**2013**). "Input to language: The phonetics and perception of infant-directed speech", Lang. Linguist. Compass **7**, 157–170.

Dempster, A., Laird, N., and Rubin, D. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm", J. R. Stat. Soc. Ser. B (Methodological) 1–38.

Englund, K. and Behne, D. (**2006**). "Changes in infant directed speech in the first six months", Infant Child Dev. **15**, 139–160.

Ferguson, C. (**1964**). "Baby talk in six languages", Am. Anthropol. **66**, 103–114.

Fernald, A. (**2000**). "Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition", Phonetica **57**, 242–254.

Fernald, A. and Simon, T. (**1984**). "Expanded intonation contours in mothers' speech to newborns", Dev. Psychol. **20**, 104–113.

Fernald, A., Taeschner, T., Dunn, J., Papoušek, M., de Boysson-Bardies, B., and Fukui, I. (**1989**). "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants", J. Child Lang. **16**, 477–501.

Fisher, C. and Tokura, H. (**1996**). "Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence", Child Dev. **67**, 3192–3218.

Grieser, D. and Kuhl, P. (**1988**). "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese", Dev. Psychol. **24**, 14–20.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (**2009**). "The WEKA data mining software: An update", ACM SIGKDD explorations **11**, 10–18.

Igarashi, Y., Nishikawa, K., Tanaka, K., and Mazuka, R. (**2013**). "Phonological theory informs the analysis of intonational exaggeration in Japanese infant-directed speech", J. Acoust. Soc. Am. **134**, 1283–1294.

Johnson, E., Seidl, A., and Tyler, M. (**2014**). "The edge factor in early word segmentation: Utterance-level prosody enables word form extraction by 6-month-olds", PLOS ONE **9**, e83546.

Jusczyk, P. W. (**2000**). *The discovery of spoken language* (MIT Press, Cambridge, MA), 328 pages.

Kawahara, S. (**2015**). "The phonetics of sokuon, or obstruent geminates", in *Handbook of Japanese Phonetics and Phonology*, edited by H. Kubozono, 43–78 (De Gruyter Mouton, Berlin).

Kirchhoff, K. and Schimmel, S. (**2005**). "Statistical properties of infant-directed versus adult-directed speech: insights from speech recognition", J. Acoust. Soc. Am. **117**, 2238–2246.

Kondaurova, M. and Bergeson, T. (**2011**). "The effects of age and infant hearing status on maternal use of prosodic cues for clause boundaries in speech", J. Speech Lang. Hear. Res. **54**, 740–754.

Kondaurova, M., Bergeson, T., and Dilley, L. (**2012**). "Effects of deafness on acoustic characteristics of American English tense/lax vowels in maternal speech to infants", J. Acoust. Soc. Am. **132**, 1039–1049.

Kuhl, P., Andruski, J., Chistovich, I., Chistovich, L., Kozhevnikova, E., Ryskina, V., Stolyarova, E., Sundberg, U., and Lacerda, F. (**1997**). "Cross-language analysis of phonetic units in language addressed to infants", Science **277**, 684–686.

Liu, H.-M., Tsao, F.-M., and Kuhl, P. (**2009**). "Age-related changes in acoustic modifications of Mandarin maternal speech to preverbal infants and five-year-old children: a longitudinal study", J. Child Lang. **36**, 909–922.

Ludusan, B. and Dupoux, E. (**2014**). "Towards low-resource prosodic boundary detection", in *Proc. of SLTU*, 231–237.

Ludusan, B., Gravier, G., and Dupoux, E. (**2014**). "Incorporating prosodic boundaries in unsupervised term discovery", in *Proc. of Speech Prosody*, 939–943.

Ludusan, B., Synnaeve, G., and Dupoux, E. (**2015**). "Prosodic boundary information helps unsupervised word segmentation", in *Proc. of NAACL-HLT*, 953–963.
Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. (**2002**). "X-JToBI: an extended J-ToBI for spontaneous speech", in *Proc. of INTERSPEECH*, 1545–1548.
Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., and Cristia, A. (**2015**). "Mothers speak less clearly to infants than to adults. A comprehensive test of the hyperarticulation hypothesis", Psychol. Sci. **26**, 341–347.
Mazuka, R., Igarashi, Y., and Nishikawa, K. (**2006**). "Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus (COE Workshop session 2)", IEICE Technical Report **106**, 11–15.
McMurray, B., Kovack-Lesh, K., Goodwin, D., and McEchron, W. (**2013**). "Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence?", Cognition **129**, 362–378.
Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. (**2009**). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement", Ann. Intern. Med. **151**, 264–269.
Morgan, J. and Demuth, K. (**2014**). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (Psychology Press, New York, NY), 504 pages.
Nespor, M. and Vogel, I. (**2007**). *Prosodic phonology: with a new foreword*, volume 28 of *Studies in generative grammar* (De Gruyter Mouton, Berlin), 327 pages.
Paccia-Cooper, J. and Cooper, W. (**1980**). "The processing of phrase structures in speech production", in *Perspectives on the study of speech*, edited by P. D. Eimas and J. Miller, 311–336 (Lawrence Erlbaum Associates, Hillsdale, NJ).
Payne, E., Post, B., Astruc, L., Prieto, P., and Vanrell, M. d. M. (**2010**). "A cross-linguistic study of prosodic lengthening in child-directed speech", in *Proc. of Speech Prosody*, paper 319.
Phillips, R. (**1994**). "Infant-directed speech in African-American mothers", Ph.D. thesis, University of Illinois – Urbana-Champaign.
Quinlan, R. (**1993**). *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, San Mateo, CA), 302 pages.
Rumelhart, D., Hinton, G., and Williams, R. (**1985**). "Learning internal representations by error propagation", Technical Report, DTIC Document.
Schmitz, M. (**2008**). "The perception of clauses in 6-and 8-month-old German-learning infants: Influence of pause duration and the natural pause hierarchy", Ph.D. thesis, Potsdam University.
Scott, D. (**1982**). "Duration as a cue to the perception of a phrase boundary", J. Acoust. Soc. Am. **71**, 996–1007.
Shneidman, L. and Goldin-Meadow, S. (**2012**). "Language input and acquisition in a Mayan village: how important is directed speech?", Dev. Sci. **15**, 659–673.
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (**1992**). "ToBI: a standard for labeling English prosody", in *Proc. of ICSLP*, 867–870.
Singh, L., Morgan, J. L., and Best, C. T. (**2002**). "Infants' listening preferences: Baby talk or happy talk?", Infancy **3**, 365–394.
Steinhauer, K. and Friederici, A. (**2001**). "Prosodic boundaries, comma rules, and brain responses: The closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers", J. Psycholinguist. Res. **30**, 267–295.
Stern, D., Spieker, S., Barnett, R., and MacKain, K. (**1983**). "The prosody of maternal speech: Infant age and context related changes", J. Child Lang. **10**, 1–15.
Supplementary Material (**????**). See Supplementary Material at [URL will be inserted by publisher].
Vaissière, J. (**1983**). "Language-independent prosodic features", in *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd, 53–66 (Springer, Berlin-Heidelberg).
Venditti, J. J. (**2005**). "The J_ToBI model of Japanese intonation", in *Prosodic typology: The phonology of intonation and phrasing*, edited by S.-A. Jun, 172–200 (Oxford University Press, Oxford).
Vosoughi, S. and Roy, D. (**2012**). "A longitudinal study of prosodic exaggeration in child-directed speech", in *Proc. of Speech Prosody*, 194–197.
Wang, Y., Seidl, A., and Cristia, A. (**2014**). "Acoustic-phonetic differences between infant- and adult-directed speech: the role of stress and utterance position", J. Child Lang. 1–22.
Weisleder, A. and Fernald, A. (**2013**). "Talking to children matters: Early language experience strengthens processing and builds vocabulary", Psychol. Sci. **24**, 2143–2152.
Wellmann, C., Holzgrefe, J., Truckenbrodt, H., Wartenburger, I., and Höhle, B. (**2012**). "How each prosodic boundary cue matters: Evidence from German infants", Front. Psychol. **3**, 580.
Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., and Amano, S. (**2007**). "Infant-directed speech supports phonetic category learning in English and Japanese", Cognition **103**, 147–162.