Bogdan Ludusan and Emmanuel Dupoux: JASA Express Letters [http://dx.doi.org/10.1121/1.4954652] Published Online 1 July 2016

The role of prosodic boundaries in word discovery: Evidence from a computational model

Bogdan Ludusan and Emmanuel Dupoux
Laboratoire de Sciences Cognitives et Psycholinguistique, École des Hautes Études en Sciences Sociales/École Normale Supérieure–Paris Sciences et Lettres Research University/ Centre National de la Recherche Scientifique, 29 rue d'Ulm, 75005 Paris, France
bogdan.ludusan@ens.fr, emmanuel.dupoux@gmail.com

Abstract: This study aims to quantify the role of prosodic boundaries in early language acquisition using a computational modeling approach. A spoken term discovery system that models early word learning was used with and without a prosodic component on speech corpora of English, Spanish, and Japanese. The results showed that prosodic information induces a consistent improvement both in the alignment of the terms to actual word boundaries and in the phonemic homogeneity of the discovered clusters of terms. This benefit was found also when automatically discovered prosodic boundaries were used, boundaries which did not perfectly match the linguistically defined ones.

Date Received: October 9, 2015 **Date Accepted:** April 4, 2016

1. Introduction

One of the necessary steps to acquire words is to segment the continuous speech signal into discrete units. Infants have an early disposition for segmenting speech at the level of large prosodic constituents. Newborns can discriminate bisyllables that contain a phonological phrase boundary from those that do not (Christophe *et al.*, 2001). By 9 months, infants determine whether or not breaks inserted in running speech respect prosodic boundaries (Jusczyk *et al.*, 1992) and between 10- and 13-months they do not recognize as words speech fragments that contain a prosodic boundary (Gout *et al.*, 2004). From this set of data, it has been proposed that infants use phrasal prosody to constrain word segmentation (Christophe *et al.*, 2003; Seidl and Johnson, 2006). Since all these results were obtained with linguistic materials carefully constructed for experimental work, we would like to investigate how useful prosodic cues are for word segmentation in more naturalistic speech corpora.

We address this question through a quantitative approach: we construct a computational model using two speech technology systems, one that detects prosodic boundaries and one that discovers word-like motifs from continuous speech, and we apply this model to running speech. Our model does not claim to be psychologically realistic, although it incorporates psychologically plausible assumptions (limited short term memory, exemplar based lexicon, use of information plausibly available to young infants). By turning on or off the prosodic component, we propose to establish a quantitative measure of how much prosody can, in principle, help word segmentation in a realistic corpus. To test for the generalizability of the effects, we analyze corpora of three languages with different phonological properties: English, Spanish, and Japanese. Finally, we discuss the implications for infant language acquisition research.

2. Model and predictions

For the spoken term discovery component, we use MODIS (Catanese *et al.*, 2013), which among the state-of-the-art systems (Ludusan *et al.*, 2014), is the most psychologically realistic. The algorithm looks for acoustic repetitions within a short time buffer, representing the short-term memory. When matching speech segments are found, they are stored in a library of acoustic token classes (motifs). This system therefore builds incrementally an exemplar-based lexicon, which is used to further parse the input. The prosodic component builds on previous work on the automatic detection of prosodic breaks (Ludusan and Dupoux, 2014). It uses only acoustic cues which are supposed to be available to infants (pitch, duration), and combines them through unsupervised clustering. The postulated prosodic breaks influence word discovery through a mechanism inspired by empirical work showing that infants do not recognize words that straddle a boundary, and therefore essentially interpret prosodic phrase boundaries as words boundaries (Gout *et al.*, 2004). This was implemented in our model with a rule truncating word candidates whenever they straddle a prosodic boundary.

The model is evaluated in terms of how well it locates word boundaries, and whether it constructs classes that are phonemically homogeneous (see Sec. 3.3). We expect that prosodic boundaries, even if only partially correct, should enhance the finding of word boundaries. This should be true in each language tested since prosodic phrase boundaries generally align to word-level boundaries. It is also possible that prosodic boundaries may help phonemic homogeneity, since it could modify the number of potential lexical matches or their variability. We run the model using automatically derived boundaries (auto), and compare the results with those obtained without any boundary (base), and with hand annotations of intonational boundaries (IPh) and hand annotations for intonational and phonological boundaries (IPh&PPh).

3. Methods

3.1 Spoken term discovery

The MODIS term discovery system (Catanese *et al.*, 2013) works by first extracting a stretch of speech in a *short term memory buffer*. At the extremity of this buffer, a small speech fragment (250 ms), called *seed*, is selected. The seed is compared against all of the stored lexical items (called *motifs*) and if no match is found, it is compared with the entire content of a short term memory buffer. Since realizations of the same word may differ in length, matching is performed taking into account all possible stretched and contractions of the time axes of the matched tokens, a technique called dynamic time warping (DTW). This results in an acoustic similarity score for each match candidate. Whenever the score is below a *similarity threshold*, before accepting the match, longer seeds are also tested (by incrementally adding more acoustic frames) as long as the recomputed score remains below the threshold. In this process, only one best matching token is retained. Seeds that do not reach a certain *critical length* are discarded. In the lexicon, each motif is a class of acoustic tokens represented by a medoid (the token closest, on average, to the other ones), and only the medoid is used to compute the similarity score with the seeds. The library is updated by adding a token to the best matching motif, or by constructing a novel motif. Finally, the input window is shifted by the duration of the seed, and a new cycle begins. After the entire dataset is processed, a post-processing step merges all tokens of the same class that overlap in time. Prosodic boundaries impact the matching step: a seed cannot be extended over a prosodic break, nor can the matching token in the buffer.

The algorithm has several important parameters to set: the *critical length*, the *buffer size* (short term memory size), and the *similarity threshold* $\varepsilon_{\mathrm{DTW}}$. We set the critical length to 0.3 s in order to match at least one syllable, and the buffer size to 30 s. The similarity threshold yields a major trade-off: for a low threshold, the algorithm finds a small number of highly homogeneous motifs, while for a high threshold, it identifies a large number of heterogeneous ones. For the experiments conducted here, the $\varepsilon_{\mathrm{DTW}}$ threshold was varied between 2.0 and 3.0 to cover a large range of variation in token heterogeneity. We used standard mel frequency cepstral coefficients (MFCCs) as input features to the algorithm: for every 25 ms frame, we computed 12 MFCCs plus the energy of the signal, along with their difference and acceleration values, every 10 ms. The DTW search used the Euclidean distance between feature vectors.

3.2 Automatic prosodic boundary detection

In order to model as closely as possible the process of early language acquisition, we chose an unsupervised method of prosodic boundary detection that uses only acoustic features. These were computed at the syllable level, since the syllable is regarded as the natural unit of speech segmentation and perception (Bertoncini and Mehler, 1981) and they include: the length of the pause following the syllable, the syllable nucleus length, the distance between the onset of the current and that of the following syllable nucleus, and the difference in fundamental frequency between the end of the current syllable and the beginning of the following one. The choice for the four acoustic cues considered was motivated by findings of perception studies in infants and adults [see Fletcher (2010) for a review of their (quasi-)universal role in the marking of prosodic boundaries] and they have been used successfully for the automatic detection of prosodic boundaries (Ludusan and Dupoux, 2014). The extracted features were then given to a clustering algorithm based on the expectation-maximization (EM) principle which returned, for each syllable, the boundary/non-boundary decision. The EM clusterer models the features by means of mixture of Gaussians and the implementation used here was the one given by the Weka toolbox (Hall *et al.*, 2009). The algorithm employs diagonal covariance matrices and was run for a maximum of 100 iterations or until

the minimum improvement in log likelihood was below $1e^{-6}$, whichever was first. The number of clusters was set to 2 and the duration features used were obtained from the annotation supplied with the corpora.

3.3 Evaluation method

The evaluation of the motifs was performed using a previously defined set of metrics (Ludusan *et al.*, 2014) by comparing them to the gold standard (boundaries and phonemes). Here we focus on two sets: the first one measures the accuracy of the segmentation process. Each discovered term is represented by a pair of boundaries and we compared the set of discovered boundaries to the reference word boundaries, defining as a match boundaries found within 30 ms of one another. We computed precision (proportion of found boundaries that match a reference one) and recall (proportion of reference boundaries that are found). The second set of metrics characterizes the goodness of matching and clustering (in MODIS, these two processes are done simultaneously and cannot be disentangled): the normalized edit distance (NED) and the coverage. In order to compute them, each found term was translated into its corresponding sequence of phonemes. The NED is defined as the per-term class average of the Levenshtein distance between each pair of strings corresponding to terms of the same class, divided by the maximum length between the two strings (e.g., "cat" and "rat" have a NED of 0.33). It can be seen as the percentage of phonemes that differ between the two strings, expressing the goodness of the matching/clustering process. The coverage represents the percentage of phonemes belonging to the discovered terms, out of the total number of phonemes in the entire corpus, characterizing the amount of repeating patterns being found. The evaluation of the automatically detected prosodic boundaries was done by comparing them to human annotated prosodic boundaries, using precision, recall and F-score metrics.

4. Materials

The materials used in this study include recordings from three languages: English, Spanish, and Japanese, which are a part of the Boston University radio news corpus (BU) (Ostendorf *et al.*, 1996), the GLISSANDO corpus (Garrido *et al.*, 2013), and the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003), respectively. The three datasets were all annotated at segmental level as well as for prosodic structure: the English data for ToBI breaks, the Spanish recordings for minor and major phrase boundaries, and the Japanese data using the X-JToBI labeling scheme.

We considered as prosodic boundaries, breaks level 3 and 4 (English), both types of coded boundaries (Spanish), and breaks level 2 and 3 (Japanese), corresponding, respectively, to phonological (PPh) and intonational phrase (IPh) boundaries in the prosodic hierarchy (Nespor and Vogel, 2007). The average number of syllables found in our data, between two consecutive phonological/intonational boundaries, was 5.8/8.6 for English, 7.6/13.4 for Spanish and 4.6/9.3 for Japanese. A summary of the three datasets is presented in Table 1.

5. Results

5.1 Automatic prosodic boundary detection

The automatic boundaries were obtained using the algorithm presented in Sec. 3.2. The overall performance (F-score/precision/recall) of the system was (0.575/.755/.465) for English, (0.553/.724/.448) for Spanish, and (0.274/.413/.205) for Japanese. In all the languages, precision was higher than recall, which indicates that the automatic system was conservative (it preferred misses to false positives), although being far from perfect, especially in Japanese.

5.2 Spoken term discovery

We illustrate the results obtained with the following systems: *base* representing the standard MODIS without any prosodic information and *auto* representing the modified

Table 1. Summary of the materials used in the experiments.

Language	Style	# speakers (F + M)	Duration	#PPh	#IPh
English	broadcast news	6 (3 + 3)	3 h	2772	5765
Spanish	broadcast news	6 (3 + 3)	3 h 25 min	4303	5619
Japanese	academic speech	8 (4 + 4)	3 h 20 min	6675	6600

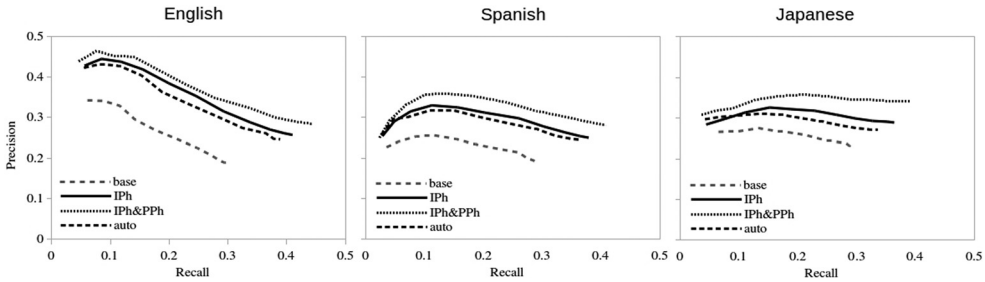

Fig. 1. Word boundary precision – recall curves obtained on the three investigated languages, across different values of ε_{DTW} (2.0–3.0), for the baseline (base) and the models incorporating intonational (IPh), intonational and phonological (IPh&PPh) or automatically obtained (auto) prosodic boundary information. The perfect model would have a precision and a recall equal to 1.

MODIS with automatically obtained prosodic boundaries. For comparison purposes, we also provide results for *IPh* and *IPh&PPh* which use gold standard IPh boundaries, or gold standard IPh and PPh boundaries, respectively. Figure 1 provides the results for segmentation quality: as expected, both boundary recall and boundary precision are higher when we introduce knowledge of prosodic boundaries. Figure 2 illustrates the trade-off between the number and the quality of the discovered motifs when varying the ε_{DTW} threshold (low coverage and low NED versus high coverage and high NED). One can see that the systems with prosody outperform the *base* system (lower NED meaning higher similarity), in all three languages, with a stronger effect for English and Spanish.

A similar experiment was performed, using more psychologically plausible features than MFCCs (compressed mel-scale magnitude features) and the same effect of prosodic boundaries was observed on the discovered terms: the phonemic similarity of the elements of a motif increased as well as the word boundary quality, albeit the overall performance was lower than in the case of MFCCs.

6. Discussion and conclusions

We have used a computational model of word discovery, inspired by studies on early language acquisition, and based on the conjunction of an automatic prosodic boundary detector and a spoken term discovery system, in order to quantitatively test the advantage of using prosodic information in word discovery. We showed that, in three languages, gold standard prosodic boundaries help in two ways: in terms of word segmentation, prosodic boundaries help to obtain more precise and more numerous word boundaries. In terms of phonemic homogeneity, prosodic boundaries help to obtain "better" word candidates (i.e., similar to one another in terms of phoneme content). Whereas the first result was expected, the second result was not necessarily obvious. We believe that the reason that homogeneity is improved by prosodic boundaries is that they provide reliable anchors for matching terms, thereby preventing the misalignment of term edges with adjacent phonemes. Indeed, an analysis of the NED across the found motifs has shown that, although there is an overall gain in NED when using prosodic information, the gain is the highest at left edge of the motif, decreasing towards the right edge. Importantly, improvements were found with automatic prosodic boundaries obtained in an unsupervised fashion. Our algorithm, which like infants, only has access to acoustic features (and not, like adults, to higher order grammatical and semantic features) discovers only around 46% of the true boundaries, and the

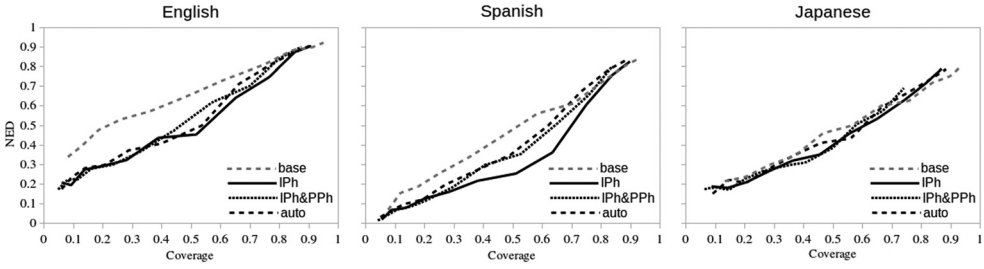

Fig. 2. Normalized edit distance (NED) – corpus coverage curves obtained on the three investigated languages, across different values of ε_{DTW} (2.0–3.0), for the baseline (base) and the models incorporating intonational (IPh), intonational and phonological (IPh&PPh) or automatically obtained (auto) prosodic boundary information. The perfect model would have a NED equal to 0 and a coverage equal to 1.

discovered boundaries are, at best, 75% correct. Despite this imperfect performance, the automatic boundaries gave a boost in performance close to the one obtained with manual IPh boundaries.

Note that beside the effects of prosody, we also uncovered differences *between* languages. In terms of word segmentation, English was found to reach the highest precision compared to Spanish and Japanese (which was further improved to a high level by prosodic information). Better segmentation of English than of Japanese has been previously found with unsupervised segmentation models working on text (phoneme sequences) and was attributed to its low segmentation ambiguity, itself due to the fact that most words tokens are monosyllabic (Fourtassi *et al.*, 2013; Ludusan *et al.*, 2015b). In the latter study, though, it was found that Japanese benefited more than English from integrating prosodic information in word discovery. Cross-linguistic differences were also found in terms of the phonemic homogeneity of discovered motifs: Japanese and Spanish have an overall higher performance than English. This effect may be driven by acoustic factors: although Japanese has difficult word minimal pairs due to the presence of long and short vowels (at least for our matching algorithm which essentially discards temporal alignment), the phonemic quality of the discovered classes was still substantially lower in English than in Japanese. We speculate that this may be due to the higher confusability of the phonemes of this language. Indirect evidence that a more crowded phonetic space could lead to higher confusability can be found in Cutler *et al.* (1996) where more missed responses were recorded in a vowel recognition task in English than in Spanish, by their respective native speakers. This may explain why English benefits most from prosodic boundary information, compared to the other two languages that already have relatively low phonemic confusability.

It remains to be seen whether these cross-linguistic differences are robust across languages and corpora, and generalizable to other speech registers, in particular, to infant directed speech (IDS) registers. Previously, we have investigated motif discovery without prosodic information in both infant and adult-directed speech (ADS), in English (Ludusan *et al.*, 2015a), and our results showed an advantage for the latter register, probably due to the higher phonetic variability present in infant-directed speech (Martin *et al.*, 2015). Based on those findings and on the current results, it would be interesting to see whether the higher incidence of prosodic boundaries in IDS would actually turn the results around (increased performance in IDS compared to ADS). We acknowledge that our analysis may be limited by the specificities of our computational components. A further validation would analyze the distribution of the motifs automatically discovered (frequency, length, phonological structure), and derive testable language-specific predictions of infant's early lexical contents [see Ngon *et al.* (2013)]. It would also be interesting to use a similar computational approach to evaluate the effect of prosody on the acquisition of other linguistic components [see Pate and Goldwater (2011) for syntactic acquisition].

Acknowledgments

The research leading to these results was funded by the European Research Council (Grant No. ERC-2011-AdG-295810 BOOTPHON). It was also supported by the Agence Nationale pour la Recherche (Grant No. ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*).

References and links

Bertoncini, J., and Mehler, J. (**1981**). "Syllables as units in infant speech perception," Infant Behav. Dev. **4**, 247–260.

Catanese, L., Souviraà-Labastie, N., Qu, B., Campion, S., Gravier, G., Vincent, E., and Bimbot, F. (**2013**). "MODIS: An audio motif discovery software," in *Proceedings of INTERSPEECH*.

Christophe, A., Gout, A., Peperkamp, S., and Morgan, J. (**2003**). "Discovering words in the continuous speech stream: The role of prosody," J. Phon. **31**(3), 585–598.

Christophe, A., Mehler, J., and Sebastián-Gallés, N. (**2001**). "Perception of prosodic boundary correlates by newborn infants," Infancy **2**(3), 385–394.

Cutler, A., Van Ooijen, B., Norris, D., and Sánchez-Casas, R. (**1996**). "Speeded detection of vowels: A cross-linguistic study," Percept. Psychophys. **58**(6), 807–822.

Fletcher, J. (**2010**). "The prosody of speech: Timing and rhythm," in *The Handbook of Phonetic Sciences*, 2nd ed., edited by W. J. Hardcastle, J. Laver, and F. E. Gibbon (Wiley, New York), pp. 521–602.

Fourtassi, A., Börschinger, B., Johnson, M., and Dupoux, E. (**2013**). "Whyisenglishsoeasytosegment," in *Proceedings of CMCL*, pp. 1–10.

Garrido, J., Escudero, D., Aguilar, L., Cardenoso, V., Rodero, E., de-la Mota, C., Gonzalez, C., Vivaracho, C., Rustullet, S., Larrea, O., Laplaza, Y., Vizcaino, F., Estebas, E., Cabrera, M., and Bonafonte, A. (**2013**). "Glissando: A corpus for multidisciplinary prosodic studies in Spanish and Catalan," Lang. Resour. Eval. **47**(4), 945–971.

Gout, A., Christophe, A., and Morgan, J. (**2004**). "Phonological phrase boundaries constrain lexical access II. Infant data," J. Mem. Lang. **51**(4), 548–567.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (**2009**). "The WEKA data mining software: An update," SIGKDD Explor. Newsl. **11**(1), 10–18.

Jusczyk, P., Kemler-Nelson, D., Hirsh-Pasek, K., Kennedy, L., Woodward, A., and Piwoz, J. (**1992**). "Perception of acoustic correlates of major phrasal units by young infants," Cogn. Psych. **24**(2), 252–293.

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

Ludusan, B., Seidl, A., Dupoux, E., and Cristia, A. (**2015a**). "Motif discovery in infant- and adult-directed speech," in *Proceedings of CogACLL*, pp. 93–102.

Ludusan, B., Synnaeve, G., and Dupoux, E. (**2015b**). "Prosodic boundary information helps unsupervised word segmentation," in *Proceedings of NAACL-HLT*, pp. 953–963.

Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.-N., Johnson, M., and Dupoux, E. (**2014**). "Bridging the gap between speech technology and natural language processing: An evaluation toolbox for term discovery systems," in *Proceedings of LREC*, pp. 560–567.

Maekawa, K. (**2003**). "Corpus of Spontaneous Japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., and Cristia, A. (**2015**). "Mothers speak less clearly to infants: A comprehensive test of the hyperarticulation hypothesis," Psychol. Sci. **26**(3), 341–347.

Nespor, M., and Vogel, I. (**2007**). *Prosodic Phonology* (Walter de Gruyter, Berlin), Vol. 28.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., and Peperkamp, S. (**2013**). "(Non) words, (non) words, (non) words: Evidence for a protolexicon during the first year of life," Dev. Sci. **16**(1), 24–34.

Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. (**1996**). Boston University radio speech corpus LDC96S36, web download (Linguistic Data Consortium, Philadelphia).

Pate, J. K., and Goldwater, S. (**2011**). "Unsupervised syntactic chunking with acoustic cues: Computational models for prosodic bootstrapping," in *Proceedings of CMCL*, pp. 20–29.

Seidl, A., and Johnson, E. (**2006**). "Infant word segmentation revisited: Edge alignment facilitates target extraction," Dev. Sci. **9**(6), 565–573.