TOWARDS LOW-RESOURCE PROSODIC BOUNDARY DETECTION

Bogdan Ludusan, Emmanuel Dupoux

LSCP - EHESS/ENS/CNRS

ABSTRACT

In this study we propose a method of prosodic boundary detection based only on acoustic cues which are easily extractable from the speech signal and without any supervision. Drawing a parallel between the process of language acquisition in babies and the speech processing techniques for under-resourced languages, we take advantage of the findings of several psycholinguistic studies relative to the cues used by babies for the identification of prosodic boundaries. Several durational and pitch cues were investigated, by themselves or in a combination, and relatively good performances were achieved. The best result obtained, a combination of all the cues, compares well against a previously proposed approach, without relying on any learning method or any lexical or syntactic cues.

***Index Terms* -** Prosodic boundaries, acoustic cues, prosody recognition

1. INTRODUCTION

Computer and mobile speech applications are being developed over an increasing number of languages, but few of these languages have labelled linguistic data in the abundant quantities necessary for the intensive training regime used in state of the art applications This creates the need to develop speech processing algorithms that operate under limited resources. At the extreme, the so-called “zero-resource” setting refers to speech tasks that are performed on the raw signal, without any use of additional expert resources (annotations) [12]. Unsupervised acoustic modelling [14] and spoken term discovery [19] are among two well known “zero-resource” algorithms that have applications in a variety of tasks: audio search, speech document classification or recommendation, etc.

Continuous speech is organized in perceptual constituents based, among others, on physiological and linguistics (semantic) factors. We define a prosodic boundary to be a delimiter of these constituents. The extraction of such type of prosodic information is not so well studied under low-resource conditions, although it has great potential in that regard, given the fact that prosody relies on relatively simple and robust cues. For instance, we know that by the middle of their first year of life, human infants are able to segment the speech input in terms of prosodic constituents [5, 6]. This ability develops in the absence of a significant lexicon or even acoustic models, and at any rate, in the total absence of annotated data.

In the literature on automatic processing of prosody of the last twenty years several boundary detection systems were proposed. Most of these systems involved some sort of learning, either in a supervised manner (e.g. [25, 10, 2]), in a semi-supervised fashion (e.g. [13]), or totally unsupervised (e.g. [1, 11, 4]). Also, they tended to rely on higher level information, like lexical or syntactic [10, 1, 2, 4, 13] or information about the structure of the studied language [11]. Among the applications for this automatically extracted information we could include automatic prosodic annotation, automatic speech recognition [10] or speech understanding systems [17]. More recently, prosodic boundaries where shown to improve spoken term discovery, by increasing the attained term precision [16].

In this paper, we explore the automatic extraction of prosodic boundaries using limited resources and low level cues. We conduct a feasibility study by exploring the validity of low level cues that have been shown to be useful to extract prosodic boundaries in human and infants. Indeed, before developing a full low-resource algorithm, it is important to determine which of these cues actually carry some weight and contribute to the recognition of prosodic boundaries in a quantitative way. Prosodic boundaries detected using a low-resource system could then be used in applications like spoken term discovery [16], which do not employ annotated data.

The paper is structured in the following way: Section 2 presents the cues investigated in this study, along with the function used to quantify them. We introduce in Section 3 the corpus used in the experiments, while in Section 4 we show the results obtained. The paper is concluded with a short discussion and some outlines of possible future directions.

2. METHOD

For the detection of prosodic boundaries we define an indicator function for each of the potential prosodic boundary cues. This function quantifies the degree to which a certain acoustic cue marks the existence of a prosodic boundary and it is normalised so that it falls between zero and one. By doing so, we can easily extend the algorithm to the case when multiple cues are considered, by simply summing the contributions of all the cues. This approach would also allow for a weighed combinations of cues, which we do not cover in this paper.

The cues considered in the current study are enumerated in the following list. We focused on these four

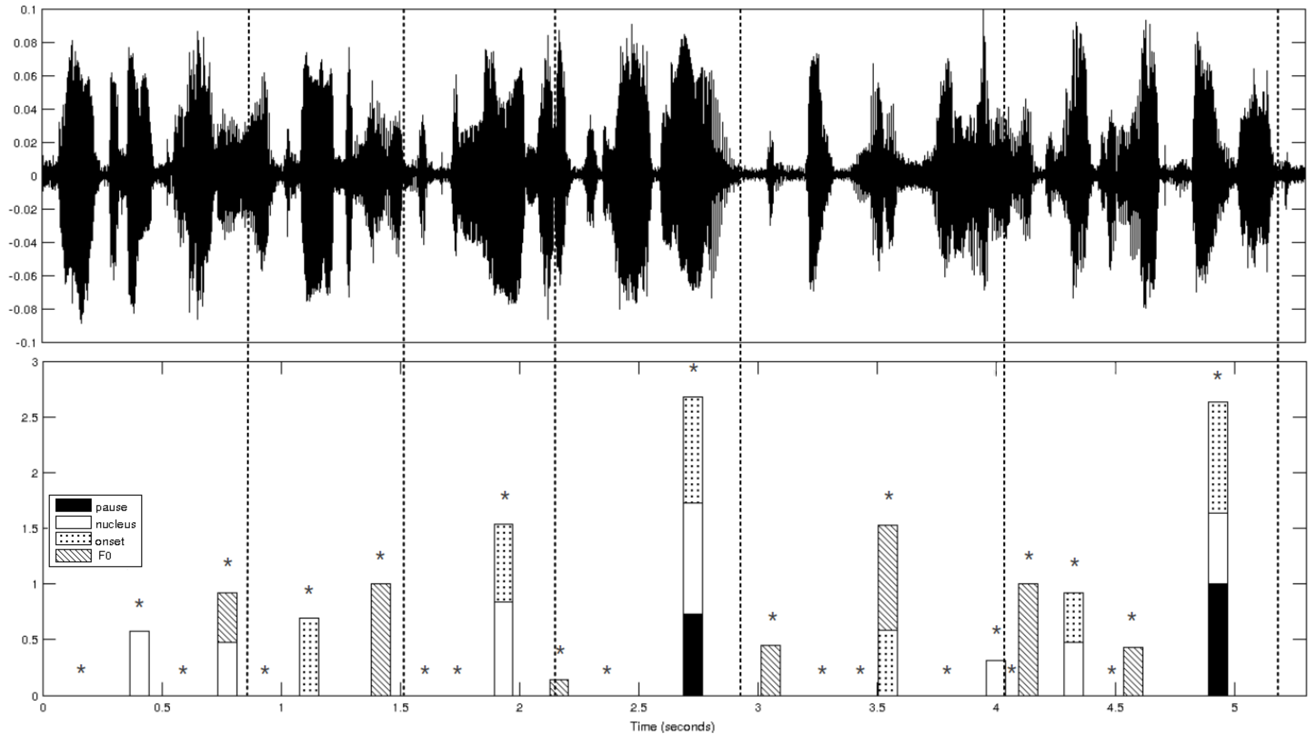

Figure 1. Waveform of a phrase from the corpus and the cumulative cues functions corresponding to it. The dashed lines in the two panels mark the position of the prosodic boundaries, while the asterisks on top of the function bars in the lower panel represent the position of each syllable nucleus.

cues because they are the most likely universal cues candidates involved in the marking of prosodic boundaries [9, 22].

- silent pauses
- nucleus duration
- nucleus-onset-to-nucleus-onset duration
- nucleus fundamental frequency (F0)

Since this paper's objective is primarily a feasibility study, we decided, for reasons of replicability, to assess the usefulness of the various features using the gold transcription. However, we restricted ourselves to cues that could, in principle, be extracted from raw speech, based on some independently published methods.

Silent pauses represent an important cue in speech perception and usually their duration is correlated to the boundary strength perceived by the listener [8]. For this reason we chose pause length as the numerical indicator of this cue, indicator which will be associated to the syllable preceding the silent pauses. There appears to be no generally agreed minimum duration for a pause, studies in the literature usually considering values between 100 and 200 ms to be the minimum [9]. Here, we set as minimum pause length 100 ms, enabling a fine grained cue which was set to zero (no pause), or to the duration of the pause. It is to be noted that we used here gold labels to indicate pauses, and thus even the shorter pauses were true pauses, not phonetic events like the closure of stop consonants.

Prosodic boundaries are also characterised by pre-boundary lengthening effects, in which the rhyme (composed of the nucleus and the coda) of the syllable immediately preceding the boundary tends to be longer than that of the same syllable not in phrase-final position. Pre-boundary lengthening is considered to be an universal phenomenon, having been investigated and observed in several dozen languages [9]. Here, we decided to use only nucleus duration because computing rhyme duration would probably require some form of alignment with an acoustic model, hence supervised learning. In contrast, nuclei can plausibly be detected using simple acoustic cues (e.g. [23, 30]). We therefore introduced as a numerical indicator of phrase-final lengthening the following function: whenever a syllable nucleus appeared to be a local maximum (i.e. with a duration more than both the preceding and following syllable), the function was set to the duration of the nucleus, otherwise it was set to zero.

As a related measure, we used the nucleus-onset-to-nucleus-onset duration (henceforth the onset cue). This cue is based on the combination of two different phenomena occurring at boundary locations: the aforementioned phrase-final lengthening and phrase-inital lengthening / strengthening. This latter phenomenon concerns chiefly the onset of the syllable just after the boundary. The measure of separation between the onsets of vowels of adjacent syllables can be therefore seen as a synthetic variable, integrating the duration of the nucleus, the duration of a potential coda, the duration of a potential pause, and the duration of the next onset. The expectation is that this measure should be maximal when the two syllables straddle a phrase boundary [7]. The corresponding numerical indicator was defined in the same manner as for the nucleus duration: it takes non-zero values only for syllables whose onset cue is a local maximum; in that case, the function will be equal to the duration of the onset cue.

Finally, we introduced a measure relative to pitch. Major phrase boundaries are usually associated with pitch resets (i.e. a large discontinuity in pitch) [22]. As we wanted to focus on this particular pattern we removed any variation due to other factors influencing the intonational pattern, by using only the mean F0 value inside the nucleus. For cases when no F0 value was found inside the nucleus (usually happening when creakiness is present) we employed an algorithm for pitch detection which returns values for any portion of the speech signal [3]. We are aware that prosodic boundaries are typically associated also with other pitch patterns. However, such patterns are typically language specific and would therefore require some learning in order to be applicable to any corpus. In the case of tonal languages, pitch fulfils additional roles, thus pitch variations could also signal lexical and grammatical identities. Still, previous studies have shown that even for tonal languages the cue chosen here, the pitch reset, is used for marking prosodic boundaries (e.g. [27]). The following indicator function was used for F0: for every nucleus corresponding to a minimum in the F0 function, we considered as measurement the size of the reset, i.e. the F0 difference between itself and the following nucleus.

Each cue function was then rescaled to 1, through a division by its maximum value over the entire news fragment file. In this way, each cue has the same importance in the calculation of the final function, when several cues are combined. Once the detection function if computed, its local maxima are determined and prosodic boundaries are placed after the nuclei corresponding to these maxima.

Figure 1 shows as an example the waveform of the phase: “My tape machine records well, but the knobs are too small, the buttons are flimsy and the counter misplaced.”, along with a cumulative representation of the acoustic cues functions computed for it. It can be observed that all boundaries are marked by at least one acoustic cue, if not several of them.

3. MATERIALS

The Boston University (BU) radio news corpus [18] was used for the evaluation of the proposed prosodic segmentation cues. We have decided to run the experiments on English because there are only a few languages which have speech resources annotated for prosody. Also, the BU corpus has been widely used for the evaluation of systems for automatic labelling of prosodic events and, thus, we can compare our results with those obtained by previously proposed algorithms. Even though English is not an under-resourced language, we believe that the cues we employ in this study are universal enough [9, 22] to be successfully used for any other language.

The BU corpus contains 10 hours of radio broadcast news recorded by 7 speakers. Out of these, approximately 3.5 hours (6 speakers) have been prosodically annotated based on the ToBI standard [21]. ToBI has 5 levels of annotation for prosodic breaks, ranging from level 0 (the least level of disjuncture) to level 4 (highest level of disjuncture). In this paper we aim at detecting intermediate phrase boundaries (ToBI level 3) and intonational phrase boundaries (ToBI level 4). We have selected for the experiments all the recordings containing both level 3 and level 4 breaks and having been aligned phonemically, giving in total approximately 3h of data. The level 3 and 4 prosodic breaks of these files were manually checked for correctness and the phoneme boundaries re-aligned by forced alignment, using HTK [29].

4. EXPERIMENTS

Several experiments were carried out in order to evaluate the importance of each individual cue, as well as that of the combination of all acoustic cues. For evaluation purposes, the boundaries were aligned to the syllable boundary following the corresponding syllable nucleus. The syllable boundaries were obtained by applying sonority-based syllabification rules to the phonemic transcription.

We evaluated the performance of the system by computing the precision, recall and F-score. The precision-recall curve of the system was obtained by varying the threshold over which a boundary decision is made. For each cue and the combination of cues, we tested 100 threshold values using the percentiles of the indicator functions.

4.1. The role of nucleus normalisation

Before examining the goodness of each individual cue, we first investigated the issue of normalisation of the nucleus duration. In general, when dealing with nuclei durations, a per-class normalisation step is applied to the data [26]. While, usually, we have access to the phoneme class, for under-resourced languages that might not be the case. Thus, this complex normalisation would be impossible to perform.

Besides the segment class, an important source of variation can be the stressed/unstressed distinction, this being especially true in the vicinity of prosodic boundaries [28]. In order to account for this difference, we first computed the ratio between the mean nucleus duration of all stressed nuclei in our corpus and that of all unstressed nuclei. The value obtained, 1.413, was then used for the normalisation of all stressed nuclei. For this study, we had access to information regarding the lexical stress from the annotation of the corpus, but even in a low-resource setting

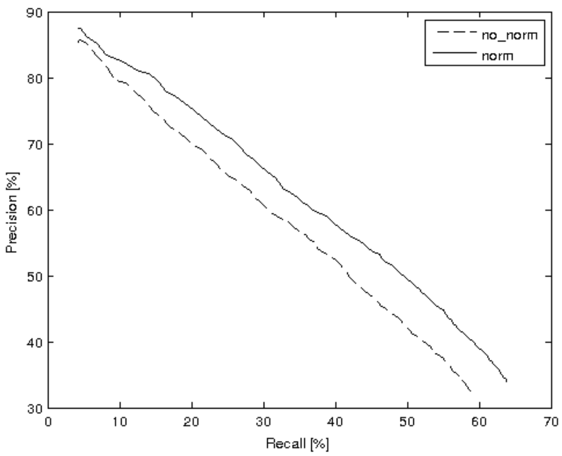

Figure 2. Precision-recall curve for non-normalised and normalised nucleus duration.

it can still be obtained, through the use of unsupervised methods of prominence detection (e.g. [15]).

The results obtained for nucleus duration are illustrated in Figure 2. It shows that even a simple stressed/unstressed normalisation method can improve performance. For the rest of the paper, when referring to nucleus duration, we intend the normalised version. Note that we do not apply normalisation to the onset cue; for it to be properly normalised, one would need a model of the duration of each of the phonemes appearing between the two onsets, something which would require supervised labels.

4.2. The performance of individual cues

The performance of each individual cue for the detection is illustrated in Figure 3. It shows a clear difference between them: a relatively low precision and a low recall for F0, higher recall and precision for the nucleus duration cue and a very high precision, but a low recall for the pause cue. The *onset* cue, which is a combination, among others, of pause and nucleus duration, gives results ranging between these last two: a similar pattern to the nucleus duration, but with a higher precision.

Table 1. Best F-score obtained for the individual cues and their combination.

Acoustic cue	*F-score*
pause	.482
nucleus	.498
onset	.519
F0	.324
sum	.588

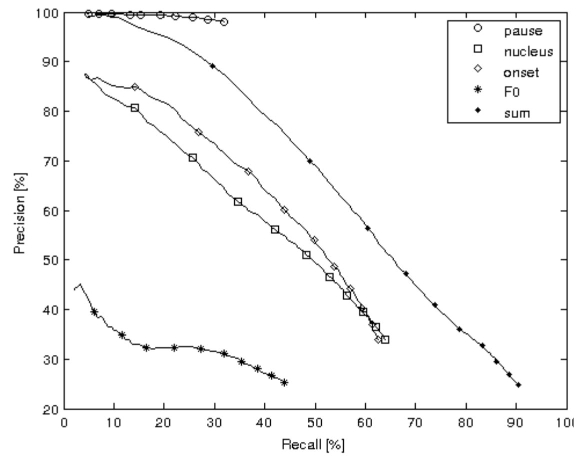

Figure 3. Precision-recall curve for each of the four acoustic cues investigated and their combination

In the first four lines of Table 1 we have displayed the best F-score obtained with each individual cue. We can see a similar distribution of the results as in the precision-recall curve, except that, due to its low recall, the *pause* results are much closer to those of *nucleus* or *onset*. These three cues are grouped around an F-score of 50%, while *F0* behaves worse, at around 30%.

4.3. The role of multiple cues

Perceptual studies have demonstrated that babies and adults can robustly perceive prosodic boundaries when more than one acoustic cue are present. Based on these findings, we explored cue combinations with the expectation that different cues could add complementary information, and/or increase the confidence of the found boundaries. We used two approaches: a simple (blind) cue combination, and a supervised optimal cue combination. The last approach is only presented to set an upper bound on what these kind of cues can achieve.

The blind cue combination is done by simply summing the corresponding individual indicator functions and applying the same criterion as for individual cues. Its performance, compared to that of individual cues can be observed in Figure 3 (*sum*).

Having used the BU corpus for the experiments, we can compare our results to previous proposals for boundary detection. While our system does not reach the same levels of performance as the proposed supervised systems [10, 2], it gives results in the same range as those obtained using an unsupervised approach, but which uses higher level information (lexical and syntactic) [1].

Ananthakrishnan's study [1] gives results for four systems, based on different clustering methods and distance measures. Their F-scores range from .58 to .66, with one system favouring precision over recall, while the rest

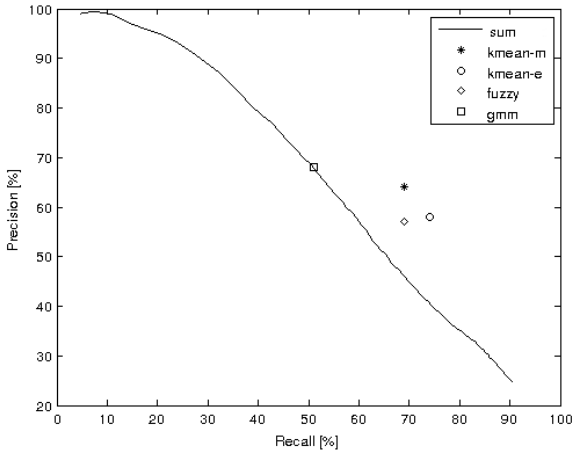

Figure 4. Comparison of the current approach with the systems proposed by Ananthakrishnan [1]

showing an opposite trend. The F-score attained with the cues combination, .588, is within this range of values. The precision-recall curve for our combination of cues is displayed in Figure 4, along with four points, representing the systems in [1]. It can be observed that the proposed approach performs the same as the most similar system (in the sense that they both favour precision over recall), *gmm*. This is true even though it employs only simple acoustic cues, without any learning or parameter optimization.

The blind cue combination that we explored above requires the sum of the indicator function to reach a certain threshold. This includes for instance a single cue with a strong value. Infant studies have shown that babies can perceive prosodic boundaries only if more than one acoustic cue would signal it [20, 24], with a similar behaviour found in adults. This suggests an additional strategy whereby the conjunction of two cues, regardless of their strength, is taken as an additional evidence for the presence of a boundary. By employing this rule, we would expect to obtain higher recall rates for the areas having a high precision.

We considered as baseline the blind combination of cues (*sum*) and we added to it different cue conjunctions. All combinations of cues were investigated, ranging from two cues to all four cues. We illustrate some of the results obtained with the conjunction of cues in Figure 5 (in the legend, P represents pause, N nucleus, O onset and F fundamental frequency). It appears that most cues combination decrease performance. The only combinations helpful are those involving two acoustic cues and one of them has a very high precision (in our case, *pause*). For them, we can observe a significant increase in terms of recall for the same level of precision, at the expense of a slightly lower maximum attainable precision. As an example, the best combination here (*PN*), has, for the maximum level of precision it obtains, a recall rate almost

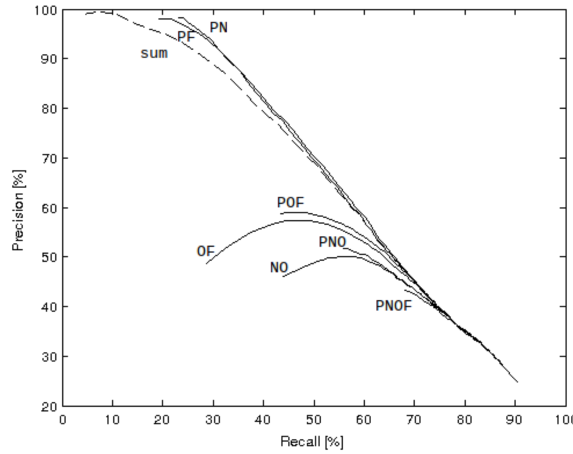

Figure 5: Precision-recall curve when multiple cues are used

double than that of the baseline (*sum*) at an equivalent precision level, at a cost of a maximum precision of 98.3% instead of 99.4%.

Although we have seen that the use of this additional strategy can help the prosodic boundary detection, it remains to be see whether the cue combination that was optimal for our corpus (*PN*) generalizes to other corpora. If optimal cue selection depends on the corpus (or language, for that reason) one could try to learn it by using as labelled boundaries those marked by silent pauses, similarly to the approach used in [11].

5. CONCLUSIONS

We have investigated in this study the usefulness of several acoustic cues in the detection of prosodic boundaries. Only cues which can be extracted directly from the speech signal were used and good results were obtained. While the best results presented here, represent the upper bound for performance using these features, our study proves that psycho-linguistically motivated acoustic cues can be used successfully for the automatic detection of prosodic boundaries. Furthermore, we have shown that combining these cues, along with considering cases of boundaries marked by multiple cues, improves performance. This is in line with findings in infant studies [20, 24] which show that babies rely on a combination of cues when deciding on the existence or not of a boundary.

We obtained good performances with these simple acoustic cues, but it was disappointing to see that the F0 cue was not very helpful. One possible cause would be the simple measurement employed in the study. In this sense, we would like to use in the future a pitch stylization algorithm in order to obtain a better estimate of the global intonational contour. A second reason might be the fact that only one F0 pattern (F0 reset) was considered here, while both intonational and intermediate phrase boundaries were sought for. We know that prosodic boundaries are marked by several types of patterns [22] and in a follow-up study we plan to investigate the learning of the F0 patterns associated to prosodic boundaries, in a similar manner to how the lexical and syntactic probabilities were estimated in [1]. By using the prosodic boundaries found through the use of the other cues, we hope to discover more F0 patterns denoting finality.

The present study used an English corpus, but we are interested in seeing how the cues used here would predict prosodic boundaries in other languages. We would expect them to be discriminative also for other languages, but probably language specific weights would have to be used. With this respect, we would like to examine how these weights can be learned with or without annotated data and if significant differences exist between these two approaches. Future work will focus on these aspects also.

6. ACKNOWLEDGEMENTS

The research leading to these results was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG) and the Fondation de France. It was also supported by ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC.

7. REFERENCES

[1] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling.", In *Proc. of INTERSPEECH-2006*, pp. 297-300, 2006.

[2] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence." *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), pp. 216-228, 2008.

[3] A. Camacho and J. Harris, "A sawtooth waveform inspired pitch estimator for speech and music." *Journal of the Acoustical Society of America*, 124, pp. 1638-1652, 2008.

[4] C. Chiang, S. Chen, H. Yu and Y. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech." J*ournal of the Acoustical Society of America*, 125, pp. 1164-1183, 2009.

[5] A. Christophe, E. Dupoux, J. Bertoncini and J. Mehler, "Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition." *Journal of the Acoustical Society of America*, 95, pp. 1570-1580, 1994.

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

[7] A. Christophe, A. Gout, S. Peperkamp and J. Morgan, "Discovering words in the continuous speech stream: The role of prosody." *Journal of Phonetics*, 31, pp. 585-598, 2003.

[8] J. de Pijper and A. Sanderman, "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues." *Journal of the Acoustical Society of America*, 96(4), 2037-2047, 1994.

[9] J. Fletcher, "The prosody of speech: timing and rhythm.", *The Handbook of Phonetic Sciences, Second Edition*, pp. 523-602, 2010.

[10] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S. Kim, A. Cohen, T. Zhang, J.Choi, H .Kim, T. Yoon and S. Chavarria, "Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus." *Speech Communication*, 46, pp. 418-439, 2005.

[11] J. Huang, M. Hasegawa-Johnson and C. Shih, "Unsupervised prosodic break detection in Mandarin speech." In *Proc. of Speech Prosody*, pp. 165-168. 2008.

[12] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, ... and S. Thomas, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition." In *Proc. of ICASSP-2013*, pp. 8111-8115, 2013.

[13] J. Jeon and Y. Liu, "Semi-supervised learning for automatic prosodic event detection using co-training algorithm." In *Proc. of the 47th Annual Meeting of the ACL*, pp. 540-548, 2009.

[14] C. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery", In *Proc. of the 50th Annual Meeting of the ACL*, pp. 40-49, 2012.

[15] B. Ludusan, A. Origlia and F. Cutugno, "On the use of the rhythmogram for automatic syllabic prominence detection.", In *Proc. of INTERSPEECH-2011*, pp. 2413-2416 , 2011.

[16] B. Ludusan, G. Gravier and E. Dupoux, "Incorporating prosodic boundaries in unsupervised term discovery.", (accepted), In *Proc. of Speech Prosody*, 2014.

[17] E. Noth, A. Batliner, A. Kießling, R. Kompe and H. Niemann. "Verbmobil: The use of prosody in the linguistic components of a speech understanding system." *IEEE Transactions on Speech and Audio Processing*, 8(5), 519-532, 2000.

[18] M. Ostendorf, P. Price and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus." *Linguistic Data Consortium*, 1995.

[19] A. Park and J. Glass, "Unsupervised pattern discovery in speech." *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), 186-197, 2008.

[20] A. Seidl, "Infants' use and weighting of prosodic cues in clause segmentation." *Journal of Memory and Language*, 57, pp. 24-48, 2007.

[21] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg, "ToBI: a standard for labeling English prosody." In *ICSLP-1992*, pp. 867-870, 1992.

[22] J. Vaissière, "Perception of Intonation." *The Handbook of Speech Perception*, pp. 236-263, 2005.

[23] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech." *IEEE Transactions on Audio, Speech and Language Processing*, 15(8), pp. 2190-2201, 2007.

[24] C. Wellmann, J. Holzgrefe, H. Truckenbrodt, I. Wartenburger and B. Höhle, "How each prosodic boundary cue matters: Evidence from German infants." *Frontiers in psychology*, 3: 580, 2012.

[25] C. Wightman and M. Ostendorf, "Automatic recognition of prosodic phrases." In *Proc. of ICASSP-91*, pp. 321-324, 1991.

[26] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf and P. Price, "Segmental durations in the vicinity of prosodic phrase

boundaries." *Journal of the Acoustical Society of America*, 91(3), pp. 1707-1717, 1992.

[27] Y. Yang and B. Wang, "Acoustic correlates of hierarchical prosodic boundary in Mandarin ." In *Proc. of Speech Prosody*, pp. 707-710. 2002.

[28] T. Yoon, J. Cole and M. Hasegawa-Johnson, "On the edge: Acoustic cues to layered prosodic domains." In *Proc. of XVIth ICPhS*, pp. 1264-1267, 2007.

[29] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "The HTK book.", *Cambridge University Engineering Department*, 3: 175, 2002.

[30] Y. Zhang and J. Glass, "Speech rhythm guided syllable nuclei detection." In *Proc. of ICASSP-2009*, pp. 3797-3800, 2009.