TOWARDS LOW-RESOURCE PROSODIC BOUNDARY DETECTION

Bogdan Ludusan, Emmanuel Dupoux

ABSTRACT

Index Terms -

1. INTRODUCTION

2. METHOD
Cues because they are the most likely universal cues 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-initial 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.

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.
4. EXPERIMENTS

4.1. The role of nucleus normalisation

3. MATERIALS
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.

<table>
<thead>
<tr>
<th>Acoustic cue</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>pause</td>
<td>.482</td>
</tr>
<tr>
<td>nucleus</td>
<td>.498</td>
</tr>
<tr>
<td>onset</td>
<td>.519</td>
</tr>
<tr>
<td>F0</td>
<td>.324</td>
</tr>
<tr>
<td>sum</td>
<td>.588</td>
</tr>
</tbody>
</table>

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...
showing an opposite trend. The F-score attained with the cues combination, 0.588, 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 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 considered.

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

sum

gmm

PN

sum

pause

PN
6. ACKNOWLEDGEMENTS

7. REFERENCES

Proc. of INTERSPEECH-2006

IEEE Transactions on Audio, Speech, and Language Processing

Society of America,

the Acoustical Society of America

Society of America

Journal of the Acoustical Society of America

Journal of Phonetics

Journal of the Acoustical Society of America

The Handbook of Phonetic Sciences, Second Edition

Speech Communication

Prosody

Proc. of ICASSP-2013

the 47th Annual Meeting of the ACL

Proc. of the 50th Annual Meeting of the ACL

Proc. of INTERSPEECH-2011

Proc. of Speech Prosody

IEEE Transactions on Speech and Audio Processing,

Linguistic Data Consortium

IEEE Transactions on Audio, Speech, and Language Processing

Journal of Memory and Language

ICSLP-1992

The Handbook of Speech Perception

IEEE Transactions on Audio, Speech and Language Processing

Frontiers in psychology

Proc. of ICASSP-91,