Is tonal alignment interpretation independent of methodology?

Caterina Petrone¹, Mariapaola D’Imperio²

¹ Zentrum für Allgemeine Sprachwissenschaft, Berlin, Germany
² Laboratoire Parole et Langage & Université de Provence, Aix-en-Provence, France
petrone@zas.gwz-berlin.de, mariapaola.dimperio@lpl-aix.fr

Abstract

Tonal target detection is a very difficult task, especially in presence of consonantal perturbations. Though different detection methods have been adopted in tonal alignment research, we still do not know which is the most reliable. In our paper, we found that such methodological choices have serious theoretical implications. Interpretation of the data strongly depends on whether tonal targets have been detected by a manual, a semi-automatic or an automatic procedure. Moreover, different segmental classes can affect target placement especially in automatic detection. This suggests the importance of keeping segmental classes separate for the purpose of statistical analysis.

Index Terms: tonal alignment, target detection methods, low tones, Neapolitan Italian.

1. Introduction

Within the autosegmental-metrical tradition, tonal targets are assumed to be temporally aligned to segmental boundaries or prosodic units, such as the syllable or the mora. Moreover, recent acoustic studies [1, 2] have suggested the existence of a strict tonal “anchoring” to segments regardless of variability factors such as segment duration or speech rate. In particular, in many languages the alignment of the L target in prenuclear rises has been found to be stably anchored to the stressed syllable onset ([1, 3, 4], inter alia). Such findings might have also important theoretical implications. So, for example, tonal target stability has been employed to support the assumption that the melodic primitives are level tones, while F0 movements are only mere interpolations [2].

Such studies often focus on differences of just a few milliseconds, thus requiring a high level of precision in target detection. However, these methodological problems have often been underestimated, and different studies have hence adopted different methods of target detection.

In many investigations, Ls and Hs are manually detected as the actual minima and maxima in the F0 contour, which are also visually identified with valleys or peaks in the pitch track. However, such targets might be difficult to detect, so that arbitrary acoustic criteria for target localization might be adopted. The L target detection in rising (or falling) pitch accents is acknowledged to be particularly problematic. In fact, the transition between the start and the end of the rise (or of the fall) is often very gradual, so that more than one possible trough can be considered as likely candidate for the L location. Moreover, segmental perturbations can either mask tonal targets (such as in presence of consonants) or create additional troughs (such as small dips at the start of the nasal consonants or drastic F0 lowering in presence of voiced obstruents). Hand-labellers can thus decide to locate the L some milliseconds (like 10 or 20 ms) before or after the perturbation, or to disregard the perturbations and take the F0 minimum at the syllable onset.

In other investigations, automatic or semi-automatic procedures of F0 modelling have been adopted, since they are claimed to be more objective than manual detection. Generally, such procedures estimate tonal targets by automatic correcting/reducing microprosodic perturbations and/or by stylizing the F0 contour. As a consequence, they are supposed to be less prone to errors due to segmental effects and also more consistent in performing target measurements.

Though different detection methods have been proposed, it is still very difficult to determine which is the most accurate, reliable and free from theoretical bias. For example, the reliability of the manual method for L target localization in American English has been supported by [5]. In this study, it has been shown that when adopting the same acoustic criteria for target detection, human labellers are more consistent than automatic procedures, especially when they are also theoretically-informed about the area in which the L is realized. However, we might still question whether “consistency” is a sufficient criterion for preferring one method over another one.

In this paper, we address the issue of whether the choice of a target detection method might also have theoretical implications. This might be especially interesting for languages where tonal alignment is exploited to convey phonological contrasts. In Neapolitan Italian, in particular, the L and H targets of the nuclear rises are later in yes/no questions (L*+H) than in narrow focus statements (L+H*) [6, 7] and this difference is realized within the stressed syllable in both intonation modalities. Also, in a perceptual study in which different F0 characteristics of the nuclear configuration (such as target alignment, scaling and rise slope) were systematically manipulated [6], it was found that alignment of the nuclear L within a LH accent can independently contribute to contour identification: the late vs. early L alignment induced a shift from question to statement interpretation. This might be taken as evidence for the phonological value of the L acoustic alignment contrast. If the methodology does not “matter”, we might expect that both manual and automatic methods would be able to capture the L alignment differences.

If, on the contrary, different detection methods affect tonal alignment results, this would also suggest that data interpretation might be biased by our methodological choices.

2. Corpus

A corpus of read speech was employed, originally recorded for other purposes [7]. Eight real words with stress on the penultimate syllable were embedded within the same carrier sentence La mamma vuole vedere la X (“(The) mum wants to see the X”, where X = target word). In such words, stressed syllable onset type as well as vowel height were orthogonally varied (Table 1). The sentences were read as either a yes/no question or a narrow focus statement, with a nuclear accent on
X and a prenuclear (L)H* on La mamma. The nuclear accent was expected to be L+H* in statements and L*+H in questions. Each sentence was repeated 7 times by 2 Neapolitan speakers (AS and DD).

<table>
<thead>
<tr>
<th>Onset type</th>
<th>/l/</th>
<th>/w/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nas</td>
<td>Mina, Nina (proper names)</td>
<td>nana</td>
</tr>
<tr>
<td>Liq</td>
<td>Lina, Rina (proper names)</td>
<td>Rana</td>
</tr>
<tr>
<td>Obs</td>
<td>Dana (proper name)</td>
<td>Dana</td>
</tr>
</tbody>
</table>

Table 1: Target words.

From a corpus of 224 utterances (2 intonation modalities X 8 words X 7 repetitions X 2 speakers), a total of 189 were analyzed. The other sentences were discarded for different reasons (inappropriate intonation, presence of creaky voice or disfluencies). Each sentence was separately stocked as .wav file. The acoustic analysis was performed through Praat [8]. We manually marked the boundaries of the target words and those of their syllables and segments. The onset (L) and offset (H) of the nuclear rise were also marked through the three procedures of tonal target detection detailed below. In this paper, only results for L target detection by consonant onset type are reported, i.e. where stronger differences among the three procedures are expected. This is motivated by two reasons. First, as seen in the Introduction, L targets are cross-linguistically more difficult to be measured in a consistent way than H targets. Moreover, in Neapolitan, nuclear Ls tend to be realized around the syllable onset [6, 7]. The vowel height manipulation is motivated only by the need of increasing the number of observations by onset type, and therefore possible effects of high vs. low vowels on L detection will be not analyzed here.

The alignment of the L target was calculated by the three methods relative to different acoustic boundaries (the onset and offset of the accented syllable, and the onset of its vowel). For the sake of brevity, we will report only results relative to consonant onset. The consonant onset has in fact been shown to be the most favorite “anchor” for alignment of L targets in accentual rises.

3. Methods

3.1. Manual Detection

The decision concerning the L localization was taken by one human labeler (the first author) by following a protocol for L detection, which is inspired to acoustic criteria for target measurement already employed in literature [1, 2, 3, 5]:

- Detect the L within the F0 region around the left periphery of the accented syllable;
- Identify the L with the F0 minimum at the start of the nuclear rise;
- If the F0 minimum is clearly due to a microprosodic effect, pick the F0 value 10 ms before the perturbation (since we assumed this type of perturbations to do not extend beyond this time span);
- If the L is realized as a low plateau, pick the F0 value at its end, i.e. when the nuclear rise begins.

In case of plateaus, many L target locations (at its start, in the middle, at its end, etc) are equally possible and we decided to arbitrarily take the end point. Note, however, that such choice is in line with results in [6], which found that in Neapolitan the end point of high plateaus corresponds to the perceived target location for the nuclear H.

The L target was first detected through a PRAAT command, which automatically found the absolute F0 minimum around the accented syllable onset for all the target words. This was done to reduce subjectivity in target detection. Such a label was automatically stocked in a point tier within the TextGrid file, and labeled as “man”. Subsequently, the waveform, spectrogram F0 contour and TextGrid of each test sentence were automatically displayed on the computer screen by a Praat script. The labeler could zoom or enlarge the edited window to better inspect the pitch contour and, in case of F0 perturbations or plateaus, to shift the point “man” found by the PRAAT command. Such changes were automatically saved in the TextGrid file.

3.2. Two-line regression (or LSF)

This semi-automatic procedure, originally developed by M. Beckman (and named by [5] “LSF”, i.e. Least Squares Fitting algorithm), has been employed, among the others, by [6]. The L target is detected at the intersection of two straight lines, fitted in a F0 region whose temporal size is pre-specified by the annotator. First, a series of piecewise linear regressions are fitted (one for each possible F0 value which is “candidate” for L target localization). In other words, each straight line is fitted to the F0 points located, respectively, at the left or the right of the F0 elbow. Then, the intersection point “selected” as L is the one of the model with the smallest total error. The success of the estimation depends on the size of the F0 region in which the algorithm looks for the elbow. In this experiment, the algorithm was fitted in the region between the start of the prosodic word and the offset of the nuclear rise. Such a choice is motivated by findings in [6,7], which showed that in Neapolitan, the nuclear L is realized close to the left periphery of the syllable, both in questions and statements. Moreover, the size of the analysis window is large enough (around 250-300 ms) to reduce segmental perturbations effects on model fitting.

The algorithm was executed in a Cygwin environment, for each test sentence. The output was saved in a .text file, and then automatically stocked in the TextGrid point tier as “lsf”.

3.3. MOMEL

The MOMEL algorithm, designed by D. Hirst and colleagues [9], is based on the hypothesis that the F0 contour is the result of the superposition of two autonomous components, the micromelody (the F0 segmental perturbations) and the macromelody (the intonation contour). The algorithm comprises four stages: After removing “aberrant” values, which are supposed to be brought about by unvoiced segments (stage 1), the resulting F0 curve is stylized by quadratic spline functions (stage 2). Such functions result into a sequence of target points (defined by their temporal location and melodic value). Each pair of targets is linked by parabolic interpolations with the spline knot at the midway point between the two targets. At this stage, relatively low F0 values are assumed to be depending on microprosodic perturbations and thus eliminated. The quadratic function is applied through a 300 ms analysis window moving along the pitch track from left to right. For each instant x, the quadratic equation provides a target point, which is located at the maximum or
minimum of the corresponding parabola. The next stage (stage 3) partitions the candidate targets and the final stage (stage 4) reduces the targets of each partition to a single candidate (see also details in [9]).

The MOMEL stylization was calculated separately for each sentence through a Praat script written by D. Hirst & R. Espesser. The result of the L detection was automatically stocked in the TextGrid point tier and labelled as “mom”.

4. Results

Fig. 1 shows the spectrogram, F0 contour and TextGrid of three test sentences, which are representative of the L localization found by the three detection procedures. We highlighted the final portion of these sentences, where the nuclear L target is expected to be realized. The figure shows the existence of a strong variation in L location according to the detection method adopted. In the upper panel, L is detected at the start of the consonant onset /d/ of “Dana” by the manual method. For the same sentence, the L is located within the F0 plateau realized before the consonant onset by the two-line regression. This can be due to the fact that the linear function underlying such a procedure could not appropriately fit the convex F0 transition between the start and the end of F0 nuclear rise realized on this token. The temporal location of the L detected by MOMEL is at the start of the article la preceding the target word.

In general, the L was generally located earlier by MOMEL than by the other two procedures. This might be partly due to an underlying difference in the notion of “target”. While, in MOMEL, the target point is defined as the focus of a parabola, both the manual detection and the two-line regression assume that the intonation contour is represented by tonal targets (i.e., turning points in the contour) linked by linear interpolation. Results from the manual and two-line detection methods are thus more similar (fig. 1; see especially the second panel). The target points found by these two methods are sometimes equal candidates for L location. In fig. 1 (lower panel), the frication noise in correspondence of /r/ brought about a local F0 perturbation. Following the protocol, the L was manually located slightly before such a perturbation. However, the two-line regression located the L after it. As a consequence, it is impossible from the mere inspection of the acoustic track to establish which is the actual target.

Figure 1: Examples of L detection for the manual procedure (“man”), the two-line regression (“LSF”) and MOMEL (“MOM”). The temporal position of the target is indicated by the circle.

In general, L was generally located earlier by MOMEL than by the other two procedures. This might be partly due to an underlying difference in the notion of “target”. While, in MOMEL, the target point is defined as the focus of a parabola, both the manual detection and the two-line regression assume that the intonation contour is represented by tonal targets (i.e., turning points in the contour) linked by linear interpolation. Results from the manual and two-line detection methods are thus more similar (fig. 1; see especially the second panel). The target points found by these two methods are sometimes equal candidates for L location. In fig. 1 (lower panel), the frication noise in correspondence of /r/ brought about a local F0 perturbation. Following the protocol, L was manually located slightly before such a perturbation. However, the two-line regression located L after it. As a consequence, it is impossible from the mere inspection of the acoustic track to establish which is the actual target.

Figure 2: Boxplots of L latency to c0 as found by the manual (upper), the two-line (“LSF”, middle) and the MOMEL (“MOM”, lower) detection methods, for both questions (“Q”) and statements (“S”). The dotted line indicates the consonant onset start.
The boxplots in fig. 2 show the values for L alignment to the start of the consonant onset (c0) for the manual (upper panel), two-line (middle) and MOMEL procedures (lower panel). Data from both speakers AS and DD are shown separately across consonantal onset type, while they are pooled across vowel height. As a first approximation, the L alignment was located later in questions than in statements by both the two-line regression (median latency value across onset types: S= -0.001 s; Q= -0.019 s for DD; S= -0.032 s; Q= 0.021 s for AS) and MOMEL (S= -0.166 s, Q= -0.062 s for DD; S= -0.174 s, Q= -0.056 s for AS). However, when the L is manually detected, the L is aligned close to c0 (S: 0.009 s; Q: 0.006 s) and AS (S and Q: 0. s.). As expected, the L is detected much earlier by MOMEL than by the two-line regression.

Fig. 2 also gives us important information about the effects of the onset type on target detection. First, the variability in target detection within the same consonant type is small for both the manual and two-line procedures, while MOMEL detection seems to be much more inconsistent. Moreover, detection results also vary across consonant onset type, though the magnitude and direction of such effects seems to partly depend on the speaker. A series of linear mixed models was therefore run separately for detection method and speakers. In such models, intonation modulation (Q/S) and onset type (Nas/Liq/Obs) were the fixed factors, while a random intercept was introduced for each of the eight words. Statistical analyses confirmed that, when the L latency from c0 is manually detected, there is no difference across intonation modality for DD (p > .01). This suggests that the L is stably anchored to the consonantal onset for that speaker. However, for AS, the L alignment is later in Q than in S with Nas (t= 2.77, p<.01) and Liq (t=3.28, p<.01), while there is no difference between the two modalities when the consonant onset is Obs. As for the two-line regression, the L was later in Q than in S, with Nas (t=3.8, p<.01) and Liq (t=1.8, p>0.04) but not with Obs onsets for DD. The Q/S effect was significant with Nas (t=7.1, p<.01), Liq (t=2.88, p<.01) and Obs (t=2.7, p<.01) for AS. Finally, when the L was detected by MOMEL, the Q/S effect was significant with Nas (t=8.4, p<.01 for DD; t=9, p<.01 for AS), Liq (t=5.4, p<.01 for DD; t=7.1, p<.01 for AS) and Obs (t=5.4, p<.01 for DD; t=4.3, p<.01 for AS) onsets for both the speakers.

5. Discussion

Through this study we have shown that methodological choice for L target detection may lead to strong variation in alignment results and, eventually, in theoretical interpretation. The variability in L detection is very small for both manual and two-line detection procedures. This is also in line with [5], which revealed that human annotators are very consistent, and that the two-line regression performs target measurements in a similar way as the manual labelling. However, our results also show that consistency is not a sufficient criterion for selecting the most appropriate detection method. As for the results obtained by the manual detection for speaker DD, the L target is aligned to the consonant onset independent of intonation modality and consonant onset type, thus supporting the segmental anchoring hypothesis. However, the manual method might dangerously obfuscate phonological information: the late alignment of the L is in fact perceptually relevant for the Q/S identification. Moreover, results for DD are partially inconsistent with those for AS, since for AS we found a significant effect of intonation modality with nasal and liquid onsets, while the L is anchored to the syllable onset only when it is constituted by an obstruct consonant.

The two-line regression is more congruent with the perceptual study in [6], showing that the independent manipulation of the L alignment affects the question/statement identification. Still, there appear to be some discrepancies across consonantal class. Similarly to the manual detection, the two-line regression found no difference in L alignment between questions and statement for obstruents. We do not know at this point whether this discrepancy should be attributed to an error on the part of the algorithm or if some other effect is in place. For example, an alternative hypothesis would be that microprosodic perturbations in presence of obstruents are so strong to “factor out” the phonological information concerning the contrast in L alignment between questions and statements. Possible articulatory or perceptual experiments could help us in finding external evidence for the actual location of the L target in presence of obstruents.

Different from the manual and two-line regression, the MOMEL algorithm is able to discern the difference between L alignment in questions and statements, across consonant classes and subjects. However, the greater variability in the temporal detection of the L target within the same consonantal class makes this method a less reliable candidate for conducting controlled alignment studies.

6. Conclusion

The choice of the methodology for tonal alignment studies can seriously affect the theoretical interpretation of the data. First, in order to be able to compare the results of alignment studies on different languages, methodologies should be comparable. We suggest that automatic procedures should be preferred to manual ones, even though more effort is required to improve automatic detection algorithms. Second, though we think that the use of different segmental material can lead to theoretical advances in understanding the interaction between tonal alignment and segments, through our study we have also shown the importance of keeping segmental class separate for the purpose of statistical analysis (against collapsing all segmental categories). In fact, consonant manner appears to affect target placement, especially in automatic detection.

7. References