COMPARING METHODS FOR LOCATING PITCH “ELBOWS”

Alex del Giudice[†], *Ryan Shosted*[‡], *Kathryn Davidson*[†], *Mohammad Salihie*[†], *Amalia Arvaniti*[†]

[†]University of California, San Diego [‡]University of Illinois, Urbana-Champaign
delgiudice@ling.ucsd.edu, rkshosted@uiuc.edu, kdavidson@ling.ucsd.edu,
msalihie@ling.ucsd.edu, amalia@ling.ucsd.edu

ABSTRACT

The labeling of “elbows” in an F0 contour is considered an enterprise beset with difficulty due to the inability of humans to locate pitch elbows with accuracy, consistency and in a manner devoid of theoretical bias. This paper investigates the extent to which human labelers agree with one another in locating elbows and how four algorithms compare to their results. Humans are found to be more consistent than has been suggested and a least-squares fitting algorithm best approaches their intuition. The success of algorithmic elbow location depends on the selection of the contour stretch in which the elbow is to be located. This elbow location is most consistent if performed by a theoretically-informed annotator, suggesting that an atheoretical annotation of F0 contours may be impossible to achieve, and ultimately undesirable.

Keywords: intonation, labeling, automatic annotation, pitch elbows

1. INTRODUCTION

The alignment of tonal targets relative to phonetic segments and structural elements like syllables and morae is a cornerstone of the autosegmental-metrical theory of intonational phonology, and generally of interest to scholars working on intonation, as most theories acknowledge a more or less close relationship between tonal targets (or the beginning and ending points of pitch movements) and the segmental structure of the speech signal (e.g. [2], [8], [9] among many). It is generally believed that the most difficult targets to measure in a consistent way are “elbows”, i.e. transitions between moving F0 and level F0 stretches (and vice versa), since such transitions are often gradual (e.g. [6]). The difficulty can be aggravated both by the smoothness of the contour and by microprosodic perturbations that can either interrupt the line of a contour or create artificial elbows, as happens when voiced consonants create short but substantial dips in F0.

Despite these difficulties, some researchers have relied on visual judgments of and comparisons between human labelers (e.g., [3], [10]) while others have used algorithmic methods (e.g. [6], [11]) to avoid inconsistency and bias.

However, to our knowledge, neither the lack of reliability in human measurements nor the robustness of algorithmic extraction has been rigorously tested. Here we compare the measurements of six human labelers to each other and to the results of four algorithms to provide some measure of reliability for all the methods tested. We recognize that there is no independent means to determine the location of a pitch elbow—such as a change in laryngeal setting—and thus reliability is more difficult to establish in these data than in data for which such independent means are available (cf. [7] where human measurements of VOT were compared with laryngeal settings in EGG signals). Because of this difficulty, we can only address the following questions:

- How variable and inconsistent are human labelers when provided with a protocol?
- To what extent do different algorithms agree with human judgments and which algorithm is the most robust in this respect?

We make no claims about the accuracy of the human labelers vis-à-vis the algorithms. When human labels cluster and the algorithm selects a proximate elbow, we claim the algorithm approaches human intuition. When human labels cluster and the algorithm selects a distant elbow, we claim the algorithm does not approach human intuition.

2. METHODS

2.1. Human annotation

Six annotators were asked to locate and label the elbow in 281 utterances from three American English corpora: polar questions and instances of “uptalk” from story reading and retelling [5], “uptalk” from a map task [4], and one-word declaratives [1]. Although the focus was on elbows involving a low level F0 stretch that turned into rising F0, we believe that our results can be generalized at least to F0 contours changing in the opposite direction (i.e. from falling F0 to a low level F0 stretch).

The human annotators were chosen because they had phonetic training but had minimal or no training in intonational phonology and therefore had no preconceived theoretical notions as to where elbows should be located in a given F0 contour. Thus the characteristics of the labelers were expected to lead to maximal variation in their measurements.

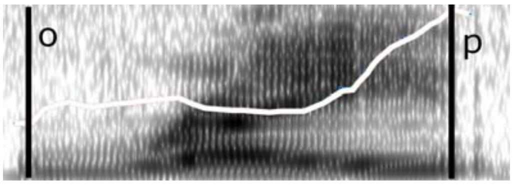
The labelers were trained by annotating ten files similar to those of the corpus together with the last author who is trained in intonational phonology. In the course of training they were advised about the effects of microprosodic perturbations on F0. These short training sessions involved several of the annotators at a time. In addition, the annotators were instructed to adhere to the following protocol:

- Focus on the pre-specified region [see below] of the intonational curve where it is believed that a transition occurs.
- Reduce the visible pitch range only insofar as the entire contour remains visible.
- Look for the point at which the highest rate of F0 change appears to occur.
- In case of a microprosodic perturbation, pick the steepest rise after the perturbation (unless it is clear that the rise occurs before it).
- If a rising line of pitch points contains a sudden increase in slope between two pitch points, ignore it if it co-occurs with a segmental boundary.

Annotation was performed using the facilities of PRAAT. The labelers used a script that allowed them to view the corpus files one at a time; they saw the waveform, spectrogram and F0 contour, along with an annotation tier in which the first author had added two labels that delimited the stretch within which the labelers were to locate the elbow. This stretch was quite broad (in all but the longest files it included practically all the speech material) and always ended at the highest F0 peak (for an example, see Fig. 1).

The location of the annotated elbows was compared across labelers by measuring for each file and annotator the distance between ‘o’ (the beginning of the region in which the elbow was to be located) and the location of the elbow itself. (Since the labelers were told to locate the elbow at some point after ‘o’, there were no negative values.) Standard deviations of the six labelers’ measurements were calculated and the distribution of the deviations was examined.

Figure 1. Elbow selection window (the waveform and annotation tier have been omitted). Labelers focused on the F0 contour between points o and p.



2.2. Algorithmic annotation

Four algorithms were used to locate the elbow in the same corpus: MAX, PER, AVG and LSF. MAX, PER and AVG are point-by-point algorithms which look at successive overlapping 3-point windows and return the first point that fulfills one of three criteria, explained below.

MAX finds the maximum pitch change across all windows; i.e.

(1) $$\max \Delta F_0$$

This algorithm is similar to the one used in [11] except that here the data were not smoothed or time-normalized.

PER finds the first window where the F0 difference exceeds a predetermined percentage of the total pitch range, i.e.

(2) $$\frac{\Delta F_0}{\max F_0 - \min F_0} > x\%$$

A 6% value was selected by trial and error. A lower setting showed sensitivity to microprosodic perturbations. A higher setting caused the algorithm to fail in files where an increase of more than 6% across three points never materialized.

AVG finds the first window where the F0 difference between its end-points exceeds the average change, i.e.

(3) $$\frac{1}{n}\sum_{i=1}^{n} \Delta F_{0i}$$

LSF, the Least-Squares Fitting Algorithm (designed by Mary Beckman and Pauline Welby, and used in [6]) operates on different principles from the other three algorithms and, crucially, operates on a larger window. This algorithm fits two lines through all the data to the left and right of each point in a pre-specified temporal interval of the F0 contour. The elbow is the intersection point of the two lines with the best fit.

3. RESULTS

3.1. Measurements by human labelers

Fig. 2 presents a histogram of the standard deviations of the six human measurements across all files. As shown, the mode was 20 ms; nearly half of the time the standard deviation was 30 ms or less, while in approximately two thirds of the data (181 files out of 281) it was 40 ms or less.

Figure 2. Histogram of standard deviations among human labelers.

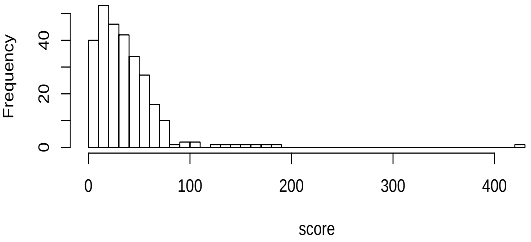
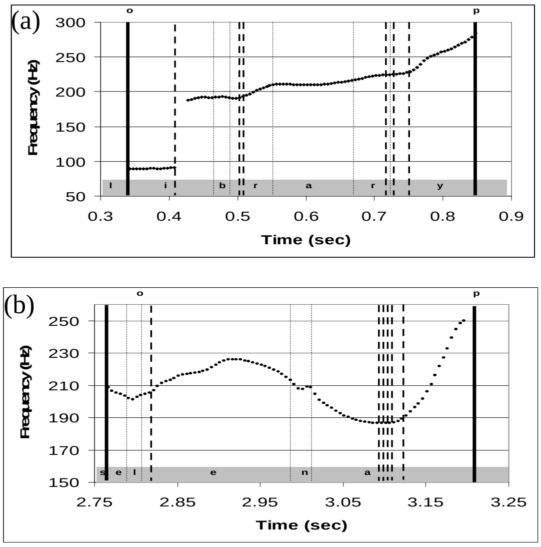


Figure 3. F0 contour of *(to the) library* in (a) and *Selena* in (b). Thin vertical lines indicate segment boundaries, dark vertical lines indicate the ‘o’ and ‘p’ boundaries, and dashed vertical lines represent labeler judgments of elbow position.

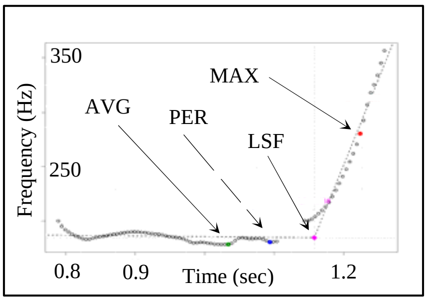


In the rest of the corpus standard deviation was 40-80 ms, and in a limited number of files it was extremely high. Inspection of several such files showed that the extreme standard deviations were due to the presence of two or more plausible elbow locations. In some cases, the two elbows were genuine in that the melody itself involved two elbows; e.g. an instance of “uptalk” autosegmentally analyzed as L* H- H% would show an elbow between L* and H- and another elbow between H- and H%, as in Fig. 3a. In other cases, however the impression of an elbow was created by a substantial F0 dip due to a voiced consonant, as in the left part of Fig. 3b. In both of these sets of data, the labelers were often divided into two groups, with one group choosing one of the possible locations, and the other group choosing the other (see Fig. 3b).

3.2. Algorithmic measurements

As shown in Fig. 4, the algorithms do not select the same point in a contour, and can vary widely in their selection. Because of the above-mentioned lack of independent evidence for the location of elbows, we used the human labels as our comparison basis. Specifically, for all comparisons we performed a deviation analysis which estimates the distance between the average of the human measurements and individual algorithmic responses to each contour. This is done by calculating the “mean elbow” for the six human labelers and then finding the difference between this point and each of the computationally-chosen elbows for each file; the average difference is then calculated across files, as shown in (4). This analysis was performed separately for the files for which the human labelers showed good agreement (n=181) and for those for which they varied widely (n=100).

Figure 4. Algorithmic results for a pitch contour.



(4)
$$\frac{\sum_{i=1}^{n}\left(\varepsilon^{i}_{\mathrm{LSF}}-\frac{\sum_{j=1}^{6}\left(\varepsilon^{i}_{\mathrm{rater}_j}\right)}{6}\right)}{n}$$

The algorithm with the smallest deviation from the averaged human elbow was LSF both in comparison with the dataset in which the humans showed good agreement, and in comparison with the set that showed disagreement (see Table 1). These results suggest that LSF most closely emulated human intuition as to where an elbow is located and could successfully handle cases of human disagreement.

Table 1: Results of deviation analysis, presented as the distance from the mean human elbow (time-normalized). Larger values indicate that the algorithm picked an elbow farther from the "mean elbow" of the six human labelers.

Labelers' s.d.	MAX	PER	AVG	LSF
s.d. < 40	18.4	21.2	25.1	8.8
s.d. > 40	23	20.9	25.7	16.6

4. DISCUSSION AND CONCLUSION

The results of the cross-labeler comparison show that the dispersion in the measurements is quite comparable to the accepted measurement error for speech segmentation, which is typically taken to be one pitch period, i.e. ±10 ms for a typical male voice. If we take into account the fact that pitch contours are more difficult to "segment" than spectrograms and waveforms, and that the labelers in the present study were not trained in intonation and were dealing with a variable corpus, not with repetitions of the same or similar utterances, this result is encouraging as it strongly suggests that human annotation is not as unreliable as it has often been taken to be.

In 36% of the data the labelers did show large differences in judgment, but these cases involved files which contained more than one possible elbow location. Thus, these extreme cases are best seen as an error due to the inexperience of the labelers and their general lack of intonation training. Errors due to microprosodic variation would not be expected to occur if the labelers had more experience with F0 contours. More importantly, the selection of one or the other of two genuine elbows would not be a problem if the labelers were guided by a theoretical understanding of what the elbows represent in each case and could focus on the same area of the overall contour. This in turn suggests that in fact having some theoretical expectations is helpful in locating pitch elbows.

This conclusion does not address the issue of bias to which theoretical expectations can lead. If bias is to be minimized, then our results suggest that the best algorithm to use is the least-squares fitting algorithm which most closely emulates human judgments, possibly because it uses a larger window and thus it is not "side-tracked" by very local events. However, the success of all algorithmic estimations also depends on selecting the right section of a contour on which the algorithm is to search for an elbow. Thus human intuition and theoretical assumptions cannot be altogether avoided. Our results show that significant deviation arises when the process is totally atheoretical, whether humans or computers do the labeling.

In conclusion, our results show that the manual annotation of elbows is not as inconsistent as it has been claimed to be, while algorithms can also be error prone. Although the use of the least-squares fitting algorithm may yield the most consistent and least biased measurements, our results show that human measurements are as reliable as other types of speech "segmentation" and that human intervention and theoretical assumptions are not only unavoidable in F0 annotation but can help constrain measurement error.

5. REFERENCES

[1] Arvaniti, A., Garding G. (in press) Dialectal variation in the rising accents of American English. In J. Cole & J. I. Hualde (eds), *Laboratory Phonology 9: Change in Phonology*. Mouton de Gruyter.

[2] Arvaniti, A., Ladd, D.R., Mennen, I. 1998. Stability of tonal alignment: The case of Greek prenuclear accents. *J. Phon.* 26, 3-25.

[3] Arvaniti, A., Ladd, D.R., Mennen I. 2006. Phonetic effects of focus and 'tonal crowding' in intonation: evidence from Greek polar questions. *Sp. Comm. 48* (6), 667-696.

[4] Barry, A., Arvaniti, A. 2006. "Uptalk" in Southern Californian and London English. Poster presented at BAAP 2006 Colloquium, Queen Margaret University College, Edinburgh, UK, April 10-12, 2006.

[5] del Giudice, A. 2006. High rising terminals in declaratives and polar question intonation of California English. Ms., University of California, San Diego.

[6] D'Imperio, M. 2000. The Role of Perception in Defining Tonal Targets and their Alignment. Ph.D. Dissertation. The Ohio State University.

[7] Francis, A.L., Ciocca, V., Yu, J.M.C. 2003. Accuracy and variability of acoustic measures of voicing onset. *J. Acoust. Soc. Am. 113* (2), 1025-1032.

[8] Ladd, D.R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[9] Ladd, D.R., Mennen, I., Schepman, A. 2000. Phonological conditioning of peak alignment in rising pitch accents in Dutch. *J. Acoust. Soc. Am. 107* (5), 2685-2696.

[10] Warren, P. 2005. Patterns of late rising in New Zealand English: Intonational variation or intonation change? *Lang. Var. Change 17*, 209-230.

[11] Xu, Yi. 1999. Effects of tone and focus on the formation and alignment of F0 contours. *J. Phon. 27*, 55-105.