PITCH ELBOW DETECTION

Uwe D. Reichel, Nele Salveste

Institute of Phonetics and Speech Processing, University of Munich
{*reichelu*|*nele*}*@phonetik.uni-muenchen.de*

Abstract: For the purpose of automatic and consistent alignment of tonal targets relative to phonetic segments we introduce one established and three new methods for automatic pitch elbow location. We further examine, whether it is beneficial to constrain the detectors to certain elbow shape types. An evaluation on hand-labeled data showed deviations from 32 to 58 ms between predicted and reference elbow locations.

1 Introduction

In the autosegmental-metrical framework [9] the alignment of tonal targets relative to phonetic segments, as for example syllable nuclei, is a key concept in intonation research (see [12, 15] for comprehensive overviews). Among the most studied tonal targets next to fundamental frequency (F0) peaks are *pitch elbows*, the transition points between moving and level pitch stretches. [8] used them to describe different types of rising intonation in Australian English. [1], and [14] examined the role of elbows in Greek and Catalan question intonation, respectively. [7] investigated elbow alignment in Neapolitan Italian. [11] analyzed the effects of syllable structure on the timing of plateaus in British English. [13] investigated the stretches of low level pitch between Dutch falling accents, and [2] applied this method for their study on rising-falling pitch accents in the Estonian variety Kihnu.

With the exception of very few studies [7, 6], in which automatic elbow detection was applied, elbows are mostly labeled manually – often with only modest inter-labeler agreement [1]. This is due to the lack of standard automatic extraction methods, which would be needed for consistent labeling, and would allow for comparing results of different studies.

As outlined in [6], pitch elbows are difficult to measure, since these transitions are often gradual and can be obscured by microprosodic perturbations. Partly based on a previous study [7] they introduced four methods for automatic elbow detection. In our study we took over the best-performing algorithm initially developed by [7] and introduce three further approaches for comparative evaluation.

2 Data

2.1 Corpus and annotation

Our data consists of 1179 Estonian single-sentence utterances by 17 speakers (ten females, seven males; age ranging from from 22 to 40 years; mean 28.2 years). Each sentence consists of two noun phrases for each of which two pitch elbows were annotated manually by a phonetic expert and Estonian native speaker (the second author). The data was recorded to examine the degrees of intonational prominence for varied focus types and utterance positions.

For the manual elbow annotation within each noun phrase, elbows were defined as the points in time before and after an F0 turning point, where F0 reaches its low or high target and does not show any further abrupt changes.

2.2 F0 preprocessing

F0 was extracted by autocorrelation (PRAAT 5.3.16 [4], sample rate 100 Hz). Voiceless utterance parts and F0 outliers were bridged by linear interpolation, and the contour was smoothed by moving median filtering with a window length of 6 samples.

3 Elbow detection methods

Within a symmetric window of 300 ms length centered on the F0 turning point within a noun phrase, the elbows left and right of the turning point were to be located automatically. The detection was carried out separately for the left respective right window half. The four methods developed for this task are introduced in the following sections.

3.1 Least-squares fit (LS)

This method first introduced in [7] performed best in the study of [6]. Two lines are iteratively fitted by means of least-squares between the syllable nucleus and the endpoint of the analysis window half. The split point of these lines is shifted sample-by-sample. For each split point the root mean squared deviation (RMS) of the two-line fit to the F0 contour is calculated. Finally, the split point for which this deviation is minimal, is chosen to be the elbow. For the current study this approach was re-implemented in terms of a piecewise linear trend removal (Matlab function *detrend*). The LS approach is illustrated in Figure 1.

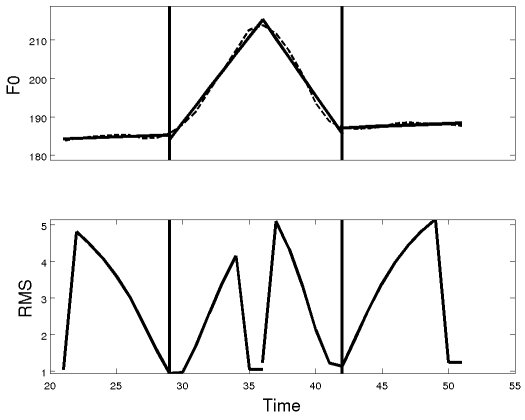

Figure 1 - Least-squares fit (LS) approach: Elbow location at split points where piecewise linear trend removal yields minimum RMS. **Top:** Best piecewise detrend analysis. **Bottom:** RMS for each split point. Detected elbows are marked by vertical lines.

3.2 Slope discontinuity (SD)

As in method LS also in method SD two lines are iteratively fitted between the syllable nucleus and the endpoint of the analysis window half. Instead of the RMS this time their slope difference is recorded. The split point at which the slope difference is highest, is considered to be the point of maximum F0 discontinuity and thus chosen as the elbow. The SD approach is illustrated in Figure 2.

3.3 Delta-Deltas (DD)

In method DD the point of maximum pitch discontinuity is determined not by linear stylization but by raw F0 values. For the pitch contour the delta-deltas of F0 values are calculated in order to identify the point of maximum F0 change discontinuity, which is then chosen to be the elbow. For concave contour shapes the maximum discontinuity is located at the the delta-delta minimum, for convex shapes at the delta-delta maximum. In section 3.5 the shape type determination is described. The DD approach is illustrated in Figure 3.

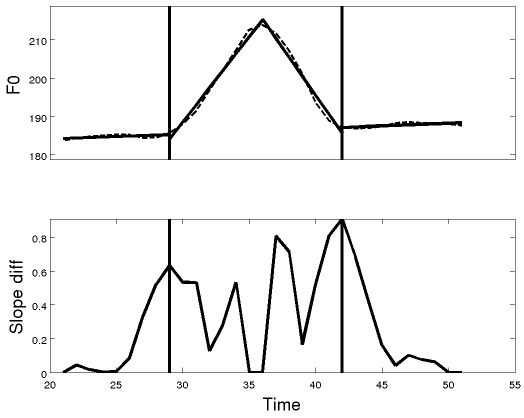

Figure 2 - Slope discontinuity (SD) approach: Elbow location at split points of maximal linear slope difference. **Top:** Best piecewise linear stylization. **Bottom:** Slope difference for each split point. Detected elbows are marked by vertical lines.

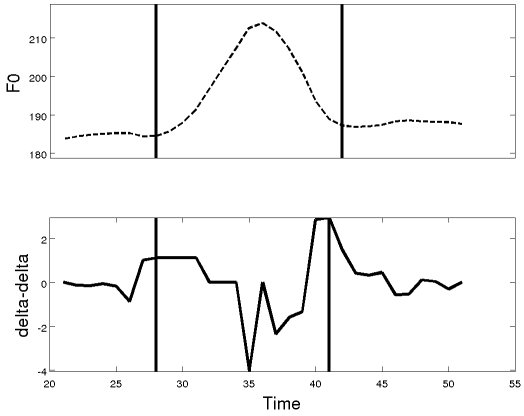

Figure 3 - Delta-Deltas (DD) approach: Elbow location at points of maximum F0 change discontinuity defined by delta-deltas. **Top:** F0 **Bottom:** Delta-delta values for each split point. Detected elbows are marked by vertical lines.

3.4 Break point detection (BP)

More generally, pitch elbow detection can be considered as a break point detection [3], i.e. to find those points in time were the parameters of a model, that underlies the data, change. For the observed F0 contour y we chose an autoregressive (AR) model. An AR model of order n and time-varying parameters ϕ is defined as:

$$y(t) = c + \sum_{i=1}^{n} \phi_i(t-i)y(t-i) + \varepsilon_t \tag{1}$$

The order n was set to 2. To estimate the model's parameters $\phi_i(t)$ changing over time we used the Kalman filter type [10] algorithm implemented in the Matlab function *segment*. It returns an $l \times n$ matrix for l time points and n model parameters. In short, m parallel models are updated at each time point t, and $\phi_i(t)$ is then derived by the weighted average of their parameters. The weights are given by the models' posterior probabilities. The model with the lowest posterior is replaced by a new model, which is derived from the most likely one by inducing a random parameter change with probability p. At the final time point l, the model with the highest posterior probability is tracked back. The time points where its parameters changed define the segment boundaries of the data.

In this study for the function *segment* default arguments were used for the number of parallel models (set to 5) and the guaranteed life time of each model (set to 1). The assumed variance of the noise $\sigma(\varepsilon_t)$ and the parameter change probability p were iteratively modified given the constraint to find (at least) one break point. The expected occurrence of one break point suggests to set the initial value for p to the inverse of the analysis window length l, $p = \frac{1}{l}$. The noise variance $\sigma(\varepsilon)$ is estimated the following way: first, the absolute y deltas are calculated, i.e. the absolute differences of subsequent F0 values. Then the noise variance is set to the variance of all deltas below the 95th percentile. This ensures, that only the most pronounced deltas above this percentile are not assumed to be noise. In case this initial restrictive setting does not yield any break point, p is successively lowered in an outer iteration loop, and σ in an inner loop until at least one break point is found, or p and σ become 1, respectively 0. The decrement stepsize for p is its initial value, and the decrement for σ is derived by successively lowering the percentile order with stepsize 5. If in the course of this iteration more than one break point has been detected the one with the greatest parameter change is kept as the elbow. This BP approach is illustrated in Figure 4.

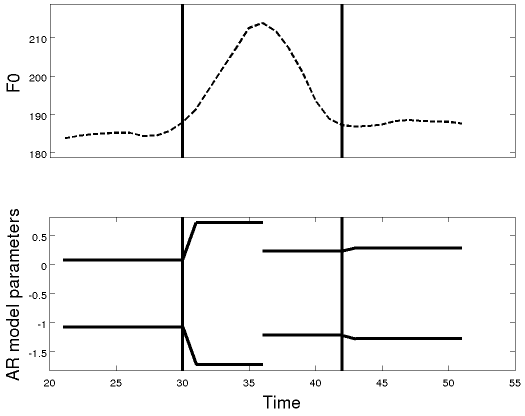

Figure 4 - Break point (BP) method: Elbow location at time points of maximal AR model parameter change. **Top:** F0. **Bottom:** Time course of piecewise second order AR model parameters. Detected elbows are marked by vertical lines.

3.5 Shape constraint (+c) and bagging (MJ)

For methods LS and SD the selection of potential split point candidates was further constrained in the following way. The overall F0 shape type left and right of the nucleus is inferred from the algebraic signs of the quadratic and linear coefficients of a second order polynomial fit. The four shape types are *concave-rising, convex-rising, concave-falling*, and *convex-falling*. For the quadratic coefficient positive values indicate convex and negative values concave shapes. For the linear coefficient, positive values indicate rising and negative values falling shapes.

Split points not in line with the overall contour shape are discarded from the list of candidates. To give an example, a convex-rising contour towards the nucleus as shown in Figure 5 suggests a flat line followed by a steep one. Split points, at which the line slopes do not fulfill these requirements, are discarded. The respective thresholds are estimated in terms of slope percentiles from the analyzed contour. Thus, only those split points are kept as elbow candidates, at which the flatter slope is smaller than the lower percentile value of all slopes, and the steeper slope exceeds the higher percentile value. If all split points have to be discarded given the initial percentile setting 10 and 90, respectively, the constraint is iteratively loosened by raising the lower percentile and lowering the upper percentile by stepsize 5 until an elbow is found.

Furthermore, the 4 methods introduced above were bagged [5] simply by taking the unweighted average of their predictions.

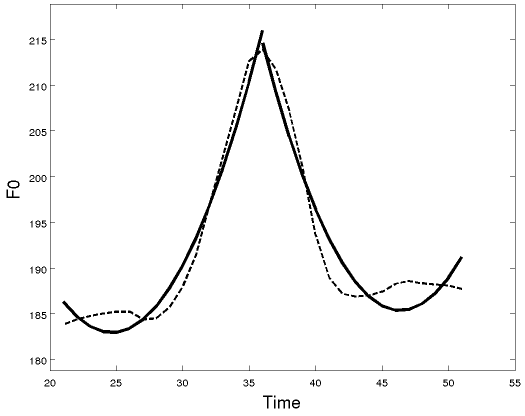

Figure 5 - Elbow shape type constraint. Shape types left and right of the F0 maximum are determined by separate quadratic polynomial fits and serve to discard split point candidates not fulfilling the shape constraint described in section 3.5. Here the left and right shape types are *convex-rising* and *convex-falling*, respectively.

4 Results

The four methods introduced in the previous section plus the application of the shape constraint were comparatively evaluated on our hand-labeled data comprising 4716 elbows. The mean deviations between predicted and reference elbow location ranged from 32 to 58 ms and are shown in Figure 6. Only one method performed significantly worse than all the others: slope discontinuity SD without shape constraint (Kruskal-Wallis test, $p < 0.001$, Dunnett post-hoc test, $\alpha = 0.05$). Thus the shape constraint $+c$ was only beneficial for the SD method. No significant performance difference was observed between the other methods except of BP and DD. The general break point detection procedure BP yielded the lowest error 32 ms and thus performed slightly better than the other approaches followed by LS with a mean error of 39 ms. Bagging all methods (MJ) did not further improve the performance but yielded a mean error of 38 ms and thus slightly worse results than BP.

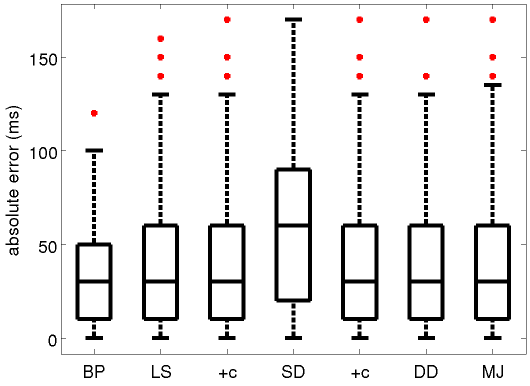

Figure 6 - Evaluation results: Mean absolute errors between elbow locations and reference positions in ms. Methods: 'BP' – break point, 'LS' – least squares, 'SD' – slope discontinuity, 'DD' – delta-deltas, 'MJ' – unweighted mean judgment of all methods. '+c' – applying the elbow shape constraint described in section 3.5 to discard unfit split point candidates.

5 Discussion

One established and three new elbow detection methods have been introduced. The best results were obtained by method BP treating elbow location more generally as a break point detection task. In the current implementation, all methods detect two elbows for each application instance, which is appropriate for our data, since it contains only cases of pairwise annotated elbows, even if the elbows are only weakly pronounced. However, the occurrence of elbows next to pitch accents is not obligatory. Except of DD all proposed methods can deal with the absence of elbows simply by switching off the iterative threshold relaxation. For methods LS and SD this refers to the shape constraint, and for method BP to the break point probability and the estimated noise variance.

One might argue, that the current evaluation results need to be treated with some caution, since our reference data is manually annotated, and manual elbow annotation had turned out to be error-prone in previous studies [1]. However, for our study the reference time stamps had been set by a phonetic expert and in accordance with well defined labeling guidelines. Thus, we consider the results to be reliable to a high extent. Since also weakly pronounced elbows had been annotated and used for evaluation, in our opinion the evaluation rather reveals lower performance boundaries. When evaluated in realistic application scenarios for detection of clearly pronounced elbows, we expect an increased performance of the methods BP, LS, and SD.

So far we did not tune our methods by parameter optimization in order to minimize the output error. In this forced elbow detection task, this optimization is not needed for the parameters mentioned in the first paragraph of this section, since they are part of the iteratively relaxed constraint definition, and thus are updated online in the course of the detection process. But optimization might be beneficial if applied to other parameters like the number of parallel models in method BP or a minimum delta-delta threshold in DD, especially if one has to cope also with the absence of elbows. Thus, the gain of optimization is still to be tested in a subsequent study.

All in all, the proposed methods allow for an automatic and consistent detection of pitch elbows and thus for comparing alignment results from different studies in intonation research.

References

[1] ARVANITI, A., D. LADD and I. MENNEN: *Phonetic effects of focus and "tonal crowding" in intonation: Evidence from Greek polar questions*. Speech Communication, 48:667–696, 2006.

[2] ASU, E. and N. SALVESTE: *The Phonetic and Phonological Analysis of the Fall-Rise Intonation Pattern in the Kihnu Variety of Estonian*. Linguistica Uralica, 3:171–179, 2012.

[3] BASSEVILLE, M. and I. NIKIFOROV: *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Englewood Cliffs, N.J., 1993.

[4] BOERSMA, P. and D. WEENINK: *PRAAT, a system for doing phonetics by computer*. Techn. Rep., Institute of Phonetic Sciences of the University of Amsterdam, 1999. 132–182.

[5] BREIMAN, L.: *Bagging predictors*. Machine Learning, 24(2):123–140, 1996.

[6] DEL GIUDICE, A., R. SHOSTED, K. DAVIDSON, M. SALIHIE and A. ARVANITI: *Comparing methods for locating pitch "elbows"*. In *Proc. ICPhS*, pp. 1117–1120, Saarbrücken, Germany, 2007.

[7] D'IMPERIO, M.: *The Role of Perception in Defining Tonal Targets and their Alignment*. PhD thesis, Ohio State University, 2000.

[8] FLETCHER, J. and D. LOAKES: *Interpreting rising intonation in Australian English*. In *Proc. Speech Prosody*, Chicago, US, 2010.

[9] GOLDSMITH, J.: *Autosegmental and metrical phonology*. Blackwell Publishers, Oxford, England, 1990.

[10] KALMAN, R.: *A New Approach to Linear Filtering and Prediction Problems*. Transaction of the ASME, J. of Basic Engineering, 82:35–45, 1960.

[11] KNIGHT, R.-A. and F. NOLAN: *The effect of pitch span on intonational plateaux*. Journal of the International Phonetic Association, 36:21–38, 2006.

[12] LADD, D.: *Intonational Phonology*. Cambridge University Press, 2009.

[13] LICKLEY, R., A. SCHEPMAN and D. LADD: *Alignment of Phrase Accent Lows in Dutch Falling Rising Questions: Theoretical and Methodological Implications*. Language and Speech, 48:157–183, 2005.

[14] PRIETO, P.: *Tonal alignment patterns in Catalan nuclear falls*. Lingua, 119(6):865–880, 2009.

[15] PRIETO, P.: *Tonal alignment*. In VAN OOSTENDORP, M., C. EWEN, B. HUME and K. RICE (eds.): *Companion to Phonology*, pp. 1185–1203. Wiley-Blackwell, 2011.