Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian

Uwe D. Reichel[1], *Katalin Mády*[2]

[1]Institute of Phonetics and Speech Processing, University of Munich
[2]Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
reichelu@phonetik.uni-muenchen,de, mady@nytud.hu

Abstract

We examined how well prosodic boundary strength can be captured by two declination stylization methods as well as by four different representations of pitch register. In the stylization proposed by Liebermann et al. (1985) base- and topline are fitted to peaks and valleys of the pitch contour, whereas in Reichel&Mády (2013) these lines are fitted to medians below and above certain pitch percentiles. From each of the stylizations four feature pools were induced representing different aspects of register discontinuity at word boundaries: discontinuities related to the base-, mid-, and topline, as well as to the range between base- and topline. Concerning stylization the median-based fitting approach turned out to be more robust with respect to declination line crossing errors and yielded base-, topline and range-related discontinuity characteristics with higher correlations to perceived boundary strength. Concerning register representation, for the peak/valley fitting approach the base- and topline patterns showed weaker correspondences to boundary strength than the other feature pools. We furthermore trained generalized linear regression models for boundary strength prediction on each feature pool. It turned out that neither the stylization method nor the register representation had a significant influence on the overall good prediction performance.
Index Terms: intonation, register, stylization, prosodic boundary

1. Introduction

The main phonetic correlates of prosodic phrase boundaries are speech pauses [1], boundary tones [2], final lowering [3], pitch reset [4], pre-final lengthening [5], and a resistance against cross-boundary coarticulation [6]. [4] have demonstrated by perception experiments with delexicalized stimuli that these acoustic features are also interpreted as boundary signals without any higher-level linguistic information.

The focus of this study lies on pitch-related discontinuity patterns at these boundaries realized as pitch reset that serves to re-initialize the pitch (F0) register to higher values after declination.

Register can be expressed in terms of *level* and *range* [7]. The level gives the distance of pitch to a reference value as for example the speaker's minimum F0. Its time course can be expressed as a base-, a mid- or a topline. The pitch range is determined by a baseline and a topline that impose a lower and upper limit for local fundamental frequency movements [8]. These lines can be calculated by means of linear regression [9] and are then defined by their F0 starting points and their slopes. In declarative sentences baseline and topline usually have negative slopes and converge towards the end of the unit, which is referred to as declination [10, 11].

As described in section 3 we generated the F0 range and the three F0 level representations in two different ways, (1) by a standard linear regression method *EXT* introduced by [9] that is still widely referred to in studies on declination, e.g. [12, 13], and a recently proposed method [14] *MED*. In *EXT* base- and topline are fitted to peaks and valleys of the pitch contour, whereas in *MED* these lines are fitted to medians below and above certain pitch percentiles. From each of the four representations we extracted a uniform set of features with the aim to grasp register discontinuities patterns at word boundaries (cf. section 4). Whereas in [14] regression trees were fitted on features across different register representations and for the stylization *MED* only, we now examined for each stylization and each feature set in isolation more systematically its relation to perceived boundary strength in terms of correlations and generalized linear regression models (cf. section 4.2).

2. Data

We analyzed 5 utterances of 10 speakers from a corpus of Hungarian spontaneous speech from map task dialogs. This corpus part is manually segmented on the word level and contains prosodic boundary labels assigned by 20 naive Hungarian subjects. The boundary label set comprises the tags *weak, strong* and *hesitation*. Hesitations and utterance-final word boundary instances were discarded for the current analysis, so that 312 word boundaries remained.

2.1. Boundary strength

In order to cope with strength judgment variation across the annotators we transformed the categorical labels into a continuous measure of perceived strength ranging from 0 to 1. For this purpose we adopted the prominence score approach of [15] expressing perceived strength as: $\frac{2 \cdot n(s)+n(w)}{2 \cdot n(\text{subjects})}$, where $n(s)$ and $n(w)$ stand for the number of *strong* and *weak* judgments respectively.

2.2. F0 Preprocessing

Voiceless segments and F0 outliers were interpolated by piecewise cubic splines. Outliers were defined as points deviating more than three standard deviations from the mean within an utterance. F0 was then smoothed by Savitzky-Golay filtering with a third order polynomial within a 5 sample window.

For speaker normalization an F0 base value b was defined as the median below the 5th percentile to be robust against non-identified outliers. F0 was then transformed to semitones (ST) relative to this base value as $F0_{st} = 12 \cdot log_2(\frac{F0_{Hz}}{b})$.

3. Stylization

At each word boundary the utterance segments of 1 second length preceding and following the boundary were taken for further analysis. The choice of 1 second is motivated by a trade-off that longer segments may contain more than one global declination event, and shorter segments may only contain local pitch events like pitch accents from which global declination cannot be inferred.

To capture F0 level and range we fitted a base-, a mid- and a topline to the F0 contour (1) within the window seg_1 preceding the boundary, (2) the window seg_2 following the boundary, and (3) within the window seg_{12} comprising both seg_1 and seg_2.

3.1. Method *EXT* for F0 level stylization

As the standard method *EXT* we employed the approach of [9] who fitted a midline through all F0 points in the respective segment, a baseline through all local valleys and a topline through all local peaks by means of linear regressions. Following [9] peaks and valleys were defined to deviate at least 10 Hz from the neighboring non-peak/valley parts of the contour. To cope with the small segment lengths the process described in the following was repeated with a successively decreased threshold until at least two peaks and valleys were detected.

- All local turning points were extracted and sorted with respect to their distance to the midline that was fitted through all data points. Weak turning points with low distance come first.
- The sorted turning points x were successively compared with the neighboring turning points in the contour. x was removed from the list of turning points if one of the two conditions held for any of its turning point neighbors y: (a) x and y were of different type (i.e. local maximum vs. minimum), and x did deviated from y by less then 10 Hz. (b) x and y were of the same type, and x was less prominent than y, i.e. a lower maximum, or a higher minimum, respectively. Note, that adjacent turning points could be of the same type in the course of the iteration, but not initially.
- The preceding step was repeated until no more local turning points were removed from the list.

Compared to a strict left-to-right processing of the local extrema this iterative approach guarantees better, that weak extrema are removed from the data first and prominent points are kept.

In order to allow for slope comparisons between seg_1, seg_2, and seg_{12} time was set to $[0\ 1]$ for both seg_1 and seg_2, and to $[0\ 2]$ for seg_{12}. Furthermore, speech pauses between seg_1 and seg_2 were removed since their length would have influenced the slope of the regression line in seg_{12} (the longer the flatter), which is not desirable in the context of the current focus on pitch discontinuity.

3.2. Method *MED* for F0 level stylization

In contrast to method *EXT*, method *MED* does not require local peak and valley detection. Here the fitting procedure consists of the following steps:

- A window of length 50 ms is shifted along the F0 contour with a step size of 10 ms.
- Within each window the F0 median is calculated
  - of the values below the 10th percentile for the baseline,
  - of the values above the 90th percentile for the topline, and
  - of all values for the midline.

  This gives 3 sequences of medians, one for the base-, the mid-, and the topline, respectively.
- For all three median sequences linear polynomials are fitted.

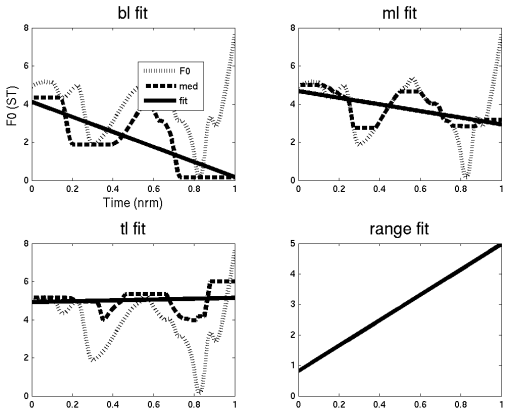

Figure 1: *Stylization of base-, mid- and topline by method* MED *based on F0 median sequences below the 10th percentile for the baseline, above the 90th percentile for the topline and for all values for the midline. The F0 range is represented by a regression line fitted through the pointwise distances between the base- and topline.*

Our method is illustrated in Figures 1 and 2. The motivation for using F0 medians relative to respective percentiles instead of local peaks and valleys is twofold. First, the stylization is less affected by prominent pitch accents and boundary tones. Second, errors resulting from incorrect local peak detection are circumvented. Both is expected to enhance stylization robustness which will be addressed in section 5.

3.3. Range stylization

Figure 1 also shows the range stylization result, that is simply derived by fitting a linear regression line through the point-wise distances between the base- and the topline. A negative slope means that base- and topline converge, whereas the positive slope in the illustrated example reflects line divergence.

4. Pitch discontinuity features

4.1. Extraction

In this study we concentrate on pitch discontinuities at word boundaries. As illustrated in Figure 3 discontinuity is measured (1) between the two segments seg_1 and seg_2 adjacent to the word boundary, and (2) between each of these segments and the joint segment seg_{12} spanning over the word boundary. (1) primarily reflects the pitch reset properties of prosodic boundaries, (2) the deviation of the pre- and post boundary F0 from a common tendency.

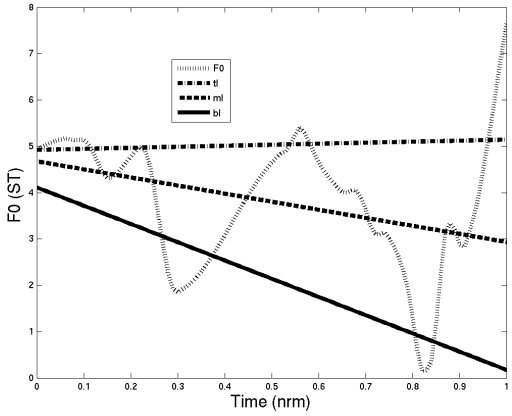

Figure 2: *Base-, mid and topline resulting from the* MED *stylization shown in Figure 1.*

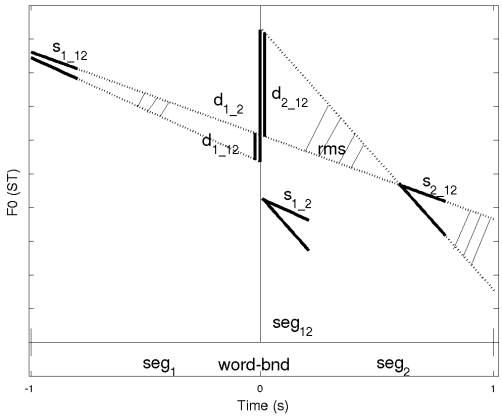

Figure 3: *Discontinuity features derived at the word boundary.* d_* *represents pitch reset properties, and* s_* *represents slope differences between the adjacent segments* seg_1 *and* seg_2*, as well as between these segments and a common declination tendency measured in* seg_{12}.

In order to ease comparisons across register representations, for both the three level and the range representation the same 7 features were extracted.

- d_{1_2}: the absolute F0 distance between the end point of the regression line in segment seg_1 and the start point of the corresponding line in segment seg_2.
- d_{1_12}: the absolute F0 distance between the end of the regression line in segment seg_1 and the corresponding time point in the line for seg_{12}.
- d_{2_12}: the absolute F0 distance between the start of the regression line in segment seg_2 and the corresponding time point in the line for seg_{12}.
- s_{1_2}: the absolute slope difference between the regression lines in seg_1 and seg_2.
- s_{1_12}: the absolute slope difference between the regression lines in seg_1 and seg_{12}.
- s_{2_12}: the absolute slope difference between the regression lines in seg_2 and seg_{12}.
- *rms*: The root mean squared distance between the concatenated lines in seg_1 and seg_2 , and the line in seg_{12}.

For low or zero-valued prosodic boundary strengths the seg_1 and seg_2 are expected to have similar declination slopes (i.e. low s_{1_2}), low pitch reset values (i.e. low d_{1_2}) and to show low deviations from a common declination tendency (low values for d_{1_12}, d_{2_12}, s_{1_12}, s_{2_12}, and *rms*). Please see Figure 3 in [14] for two examples taken from the examined corpus.

In the following the base-, mid-, topline, and range-related feature sets are referred to as *bl, ml, tl*, and *rn*, respectively.

4.2. Boundary strength prediction

For each of the two stylizations and each of the four feature sets we fitted and evaluated generalized linear models ([16]; Matlab function GeneralizedLinearModel.fit) in order to map the features introduced above to the perceived prosodic boundary strength. To restrict the output to the strength score interval [0 1] (cf. section 2) the distribution of the response was set to *binomial*, and a *logit* link function was chosen defining the relation between the linear combination of the predictors and the mean response.

5. Results

5.1. Robustness

In order to compare the robustness of the stylization methods *EXT* and *MED* we simply counted the instances in which the fitted base- and topline crossed in seg_1, seg_2, or seg_{12}, which is obviously to be regarded as an error. The error rate of *EXT* amounts 16.6% (155 out of 936 instances, that are given by 3 segments for each of the 312 word boundaries). The error rate of *MED* amounts 4.5% and thus is considerably lower.

5.2. Correlations to perceived boundary strength

Figure 4 shows the correlations of all feature sets each extracted twice from the stylization method *EXT* and *MED*.

Overall the correlations turned out to be significantly lower than the correlation between pause length and boundary strength which amounted to 0.58 for our data (two-sided one-sample sign tests for median comparison, $p < 0.05$). Whereas the median correlation was significantly higher than 0 for method *MED* for all feature sets, for method *EXT* only the feature set *ml* showed a sufficiently high correlation (two-sided one-sample sign tests for median comparison, $p < 0.05$).

We measured separately for each stylization method, whether there was a significant correlation difference dependent on the chosen feature set. For method *MED* no significant difference across the feature sets is to be reported (Kruskal-Wallis test, $p > 0.89$), for method *EXT* the features sets *bl* and *rn* were significantly lower correlated to boundary strength than *ml* (Kruskal-Wallis test, $p < 0.01$, Dunnett post-hoc test, $\alpha = 0.05$).

Finally we compared separately for each feature set, whether there was a significant difference across the stylization methods *EXT* and *MED*. *MED* yielded significantly higher correlations for *bl, tl*, and *rn* (Wilcoxon two-sided signed rank test for paired samples, $p < 0.05$), but not for *ml* ($p > 0.8$).

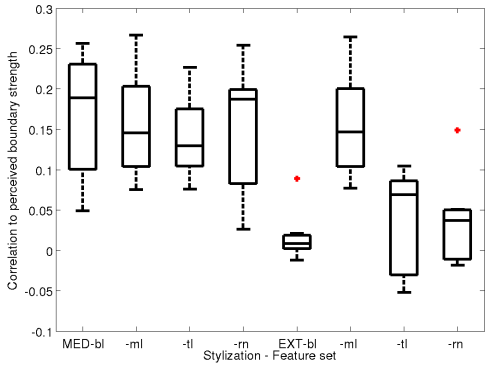

Figure 4: *Correlations of feature sets to perceived boundary strength.*

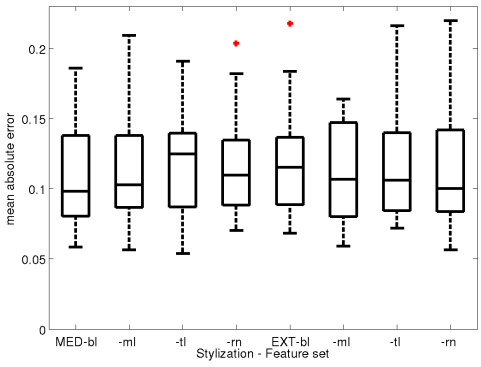

Figure 5: *Mean absolute error in 20-fold cross validation for the generalized linear regression models trained on the respective feature sets.*

5.3. Prediction of perceived boundary strength

For each stylization method and each feature set, we measured in a 20-fold cross validation the mean absolute error of the generalized linear regression model predictions and the perceived prosodic boundary strengths for the held-out data. The results are shown in Figure 5.

Overall the mean absolute errors ranged from 0.06 to 0.22 (the maximally possible error would amount 1). There aren't any significant performance differences to be reported, neither related to the stylization method nor to the feature set (ANOVA with stylization method and feature set as independent factors and the error as dependent factor, $p > 0.69$ for the method, > 0.94 for the set. Wilcoxon two-sided signed rank test for paired samples for each feature set compared between the stylization methods, $p > 0.3$).

6. Discussion and Conclusions

6.1. Declination stylization

As reported in section 5.1 method *MED* turned out to be more stable than *EXT* in avoiding implausible declination line crossings. The reason is that *MED* is less prone to two potential sources of stylization errors: first, it does not require the detection of local peaks and valleys, and second, it is less affected by local pitch events such as prominent pitch accents or boundary tones. A possibility to weaken the misleading influence of boundary tones would be to truncate the F0 contour at its ends before fitting the regression line, but it might be difficult to automatically decide when such a truncation is justifiably. Furthermore, as the current study is working on relatively short segments of 1 and 2 seconds length truncation might result in contours which are too short for a reliable declination estimation. Difficulties of the *EXT* method in finding appropriate base- and toplines is illustrated in Figure 6. It is shown that the slopes of the lines are dependent (1) on the somewhat arbitrary threshold choice for automatic peak and valley detection, and (2) on the decision whether or not to truncate contour ends.

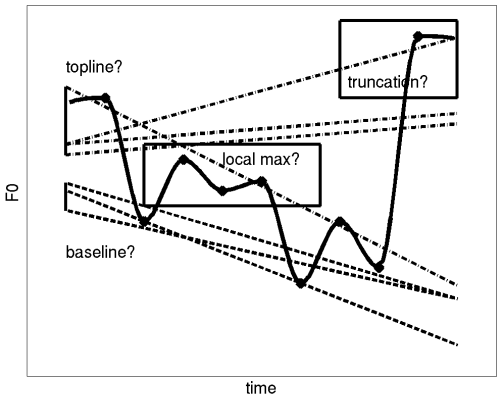

Figure 6: *Problems of register stylization on the basis of local F0 peaks and valleys: fuzzy local peak detection and high dependency of the regression result on the choice of relevant peaks and valleys leads to 3 different baselines and 4 different toplines.*

As opposed to the robustness difference in base- and topline and this range stylization, especially for relatively short segments, the midline stylization turned out to be equally robust for both methods, so that the additional median filtering step of *MED* is dispensable.

6.2. Discontinuity features and boundary strength prediction

For all feature sets the correlations to perceived prosodic boundary strength turned out to be considerably lower than for pause length reflecting the major impact of speech pauses on boundary strength. Nevertheless, it was possible to train generalized linear regression models, that were robust against different stylizations and feature sets, to predict perceived boundary strength with acceptable proximity on held-out data. These models can be of use for automatic prosodic boundary extraction, especially in cases were obvious boundary markers as speech pauses are absent.

7. Acknowledgments

The work of the first author has been carried out within the CLARIN-D project [17] (BMBF-funded). The second author was funded by OTKA 101050 "A laboratory phonology approach to Hungarian prosody".

8. References

[1] M. Swerts and R. Geluykens, "Prosody as a marker of information flow in spoken discourse," *Language and Speech*, vol. 37, no. 1, pp. 21–43, 1994.

[2] G. Brown, K. Currie, and J. Kenworthy, *Questions of Intonation*. London: Croom Helm, 1980.

[3] M. Liberman and J. Pierrehumbert, "Intonational Invariance under Changes in Pitch Range and Length," in *Language Sound Structure*, M. Aronoff and R. Oehrle, Eds. Cambridge, MA: MIT Press, 1984, pp. 157–233.

[4] J. de Pijper and A. Sandermann, "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues," *Journal of the Acoustical Society of America*, vol. 96, pp. 2037–2047, 1994.

[5] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price, "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries," *JASA*, vol. 91, no. 3, pp. 1707–1717, 1992.

[6] T. Cho, "Prosodically-conditioned strengthening and vowel-to-vowel coarticulation," *Journal of Phonetics*, vol. 32, pp. 141–176, 2004.

[7] T. Rietveld and P. Vermillion, "Cues for Perceived Pitch Register," *Phonetica*, vol. 60, pp. 261–272, 2003.

[8] K. Pike, *The intonation of American English*, ser. University of Michigan publications. Ann Arbor: University of Michigan Press, 1945, vol. 1.

[9] P. Liebermann, W. Katz, A. Jongman, R. Zimmerman, and M. Miller, "Measures of the sentence intonation of read and spontaneous speech in American English," *J. Acoust. Soc. Am.*, vol. 77, no. 2, pp. 649–657, 1985.

[10] A. Cohen, R. Collier, and J. t'Hart, "Declination: construct or intrinsic feature of speech pitch," *Phonetica*, vol. 39, pp. 254–273, 1982.

[11] D. Ladd, "Declination: A review and some hypotheses," *Phonology Yearbook*, vol. 1, pp. 53–74, 1984.

[12] M. Swerts, E. Strangert, and M. Heldner, "F0 declination in read-aloud and spontaneous speech," in *Proc. ICSLP*, vol. 3, Philadelphia, 1996, pp. 1501–1504.

[13] C. Schmid and M. Gendrot, C. Adda-Decker, "Une comparaison de déclinaison F0 entre le français et l'allemand journalistiques," in *JEP-TALN-RECITAL*, vol. 1, 2012, pp. 329–336.

[14] U. Reichel and K. Mády, "Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech," in *Proc. Elektronische Sprachverarbeitung*, ser. Studientexte zur Sprachkommunikation. TUDpress, 2013, pp. 223–23.

[15] Y. Mo, J. Cole, and M. Hasegawa-Johnson, "Prosodic effects on vowel production: evidence from formant structure," in *Proc. Eurospeech*, Brighton, 2009, pp. 2535–2538.

[16] P. McCullagh and J. Nelder, *Generalized Linear Models.* New York: Chapman&Hall, 1990.

[17] "http://eu.clarin-d.de/index.php/en/," Clarin-D web page.