F0 discontinuity as a marker of prosodic boundary strength in Lombard speech

Štefan Beňuš [1], *Uwe D. Reichel* [2], *Juraj Šimko* [3]

[1] Constantine the Philosopher University in Nitra & II SAS Bratislava, Slovakia
[2] Institute of Phonetics and Speech Processing, University of Munich, Germany
[3] Institute of Behavioral Sciences, University of Helsinki, Finland

sbenus@ukf.sk, reichelu@phonetik.uni-muenchen.de,juraj.simko@helsinki.fi

Abstract

Prosodic boundary strength (PBS) refers to the degree of disjuncture between two chunks of speech. It is affected by both linguistic and para-linguistic communicative intentions playing thus an important role in both speech generation and recognition tasks. Among several PBS signals, we focus in this paper on pitch-related discontinuities in boundaries conveying linguistically meaningful contrasts produced in increasing levels of ambient noise. We compare several measures of local and global pitch reset and use classifiers in an effort to better understand the relationship between the degree of ambient noise and F0 marking of PBS. Our results include a positive effect of some noise on boundary classification, better performance of local than global reset features, and more systematic behavior of F0 falls compared to rises.

Index Terms: prosodic boundary, Lombard speech, Slovak

1. Introduction

Speech is structured in units of the prosodic hierarchy [7,10] and this chunking plays important role in coding and decoding communicative meanings. Prosodic units are delimited by prosodic boundaries that can be characterized by their strength and type. Strength refers to the degree of disjuncture between two units flanking the boundary, and type is commonly associated with its tonal realization such as falls, rises, or plateaus for higher level boundaries. Both strength and type vary as a function of multiple linguistic and paralinguistic meanings. For example, weaker boundaries convey tighter pragmatic alignment between chunks than stronger boundaries, rises imply that there is more to follow, or plateaus might signal boredom.

Communicative meanings associated with prosodic boundaries are most commonly signaled with the degree of pre-boundary lengthening, duration of silence if present, pitch movement on the pre-boundary material, and pitch reset [13,14]. It has been shown that F0 discontinuity is a useful predictor of boundary strength both in relating observable features to theoretical frameworks [5] and in statistical modeling [8,12]. For example, a system with pitch discontinuity features in a corpus of Hungarian conversational speech [8] achieved good correlation with human judgments of boundary strength and turned out to be more robust when compared to the established fitting procedure of [6]. Furthermore, discontinuity features showed a higher correlation to perceived prosodic boundary strength when being derived from this stylization. In this work, we focus on a single boundary marker – pitch reset – and we test if the discontinuity parameterization proposed for Hungarian can be extended to classifying prosodic boundaries in other languages and different communicative domains. Particularly, we are interested in the role of both local F0 reset across a boundary as well as more global reset of declination trends [3] in cuing the strength and type of boundary.

Finally, we believe that a useful approach to better understanding the underlying system of speech production in general, and prosodic boundary realization in particular, is to employ communicatively meaningful hyper-articulation while speaking. Our assumption is that communicatively salient features participating in cuing meaningful differences both at the linguistic and paralinguistic levels will be more prominent in hyper-articulation. Since most spoken interactions occur under some ambient noise, Lombard speech is an ecological way for inducing hyper-articulation in laboratory in a controlled, yet natural, way. Lombard effect on F0 is generally linked to increased range and mean, and specifically for boundaries, to the expanded pre-boundary F0 movements and cross-boundary resets [3,12].

The core question of this paper is thus whether assumed increase in boundary strength due to increased ambient noise can assist an automatic boundary classification system based on pitch reset features. We shall also asses the dependence of several reset measures on the ambient noise level and compare the results with automatic classification performance.

2. Methodology

2.1. Corpus

Five native Slovak speakers (3F, 2M) read multiple repetitions of 12 prompt sentences with identical syllable counts (17) under 5 levels of ambient noise while both acoustic and articulatory data were collected [1,11]. Several controlled segmental manipulations of the pre-boundary rhyme and 2 post-boundary syllables were balanced and will not be analyzed here. The crucial manipulation relevant for this study is the type of the prosodic boundary between the 13th and 14th syllable in the material comprising the last 7 syllables of the prompts that are separated from the previous material by a major silent break as shown in (1) below.

(1)	#{a,ai}	xxx	C{i:,a:}{m,n}	(#)	{i,a}	{ba,bi}	xxx	xxx.
	σ11	σ12	σ13		σ14	σ15	σ16	σ17

Syllables σ12 and σ13 form a word with a lexical stress and likely also a pitch accent on the initial syllable σ12. Syllables σ14 and σ15 also form a word (*iba* 'only' and *aby* 'so that') that is typically not pitch-accented.

These two particles/conjunctions allow for a pragmatic manipulation of the prosodic boundary preceding them. Hence, using the sentence meaning and punctuation, we elicited three boundary types: **B0** with no punctuation in the prompts inducing a weak disjuncture, **B1** with a comma and inducing a major prosodic boundary realized with an F0 rise, and **B2** with a full stop and inducing a major boundary realized with an F0 fall.

In the reference "0" condition, subjects were instructed to speak normally and had no headphones. In other three conditions, subjects heard babble noise of 60, 70, 80 dB(A) through headphones in blocks. Finally, greatest hyper-articulation was assumed to arise from 80dB noise when subjects were instructed to read sentences to a non-native speaker (present and visually interacting with the subject). We refer to this condition as "80nn".

Subjects S3 and S4 produced a full intended stimuli set with 5 repetitions of each prompt within a block and 2 sets of blocks with noise conditions, S5 had a decreased number of repetitions in some blocks, S1 produced only one set of noise blocks with 5 repetitions , and S2 only 1 set with 3 repetitions. The makeup of the corpus is summarized in Table 1.

Table 1. *Corpus tokens separately for subjects and boundary types.*

Subject	B0	B1	B2	Total
S1	128	122	112	362
S2	51	57	55	163
S3	202	193	209	604
S4	195	199	200	594
S5	171	169	176	516
Total	747	740	752	2239

2.2. Features

The acoustic signal was automatically aligned with the transcription of the prompt sentences and the alignment of segments in syllables σ13-15 was hand corrected. F0 contours were extracted using an adjusted two-step pitch tracking procedure suggested in [4] and implemented in Praat.

One approach to analyzing the effect of Lombard condition on the realization of boundary strength is to treat discrete noise conditions from the elicitation (0, 60, 70, 80, 80nn) as independent variables. Additionally, we will employ a continuous proxy of the subject's response to noise based on the discrete cosine transform of the F0 extracted from the pre-boundary long vowel. This is because previous analyses of this corpus [1,11] showed that a) the order of the noise blocks significantly affected the degree of hyper-articulation (e.g. material in 60 block was more hyper-articulated when preceded by 80 than 0 block), and b) F0 mean operationalized with the first coefficient of the discrete cosine transform (*dct1*), that depicts the overall pitch level, offers a good approximation to the degree of hyper-articulation. Hence, *dct1* will serve as a continuous independent variable.

Dependent variables come primarily from F0 stylization proposed in [8]. Briefly, in pre-processing, voiceless segments and F0 outliers were linearly interpolated. Outliers were defined as points deviating more than two standard deviations from the mean within an utterance. F0 was then smoothed by Savitzky-Golay [9] filtering with a 3rd order polynomial within a 5 sample window. For speaker normalization an F0 base *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 in (2).

$$F0_{st} = 12 * \log_2(F0_{Hz} / b) \quad (2)$$

For stylization itself and using only uncorrected automatically aligned information, pre-boundary seg_1 spanned the interval between the onset of σ11 and offset of σ13 in (1), and the post-boundary seg_2 spanning from the onset of σ14 and offset of σ17 were used. To capture F0 level and range trends within these segments and within a joint segment seg_{12} spanning over seg_1 and seg_2 we fitted a base-, a mid- and a topline to the F0 contours of each segment. Time was normalized for seg_1 to the range [-1 0], for seg_2 to [0 1], and consequently time of seg_{12} to [-1 1].

For the line fits in all three segments we 1) shifted a window of length 50 ms along the F0 contour with a step size of 10 ms, 2) calculated the F0 median within each window of the values below the 10th percentile for the baseline, above the 90th percentile for the topline, and all values for the midline respectively, and 3) fitted linear polynomials for all three median sequences. The motivation for using F0 medians relative to respective percentiles instead of local peaks and valleys is twofold: the stylization is less affected by prominent pitch accents and boundary tones, and errors resulting from incorrect local peak detection are circumvented. This method is illustrated in Fig. 1 that also shows the range stylization result (the rising double line), that is derived by fitting a linear regression line through the point-wise distances between the baseline and the topline. A negative slope means that baseline and topline converge, whereas the positive slope in the illustrated example reflects line divergence.

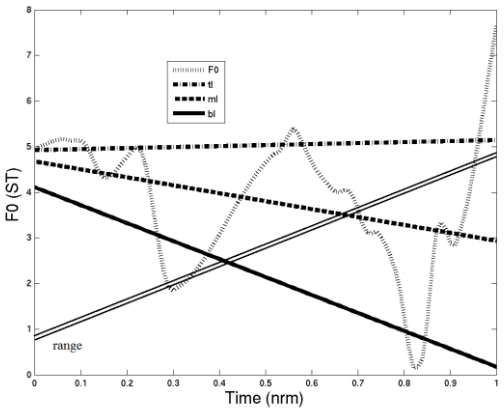

Figure 1: *Stylization of F0 contour; see text for details.*

From this stylization, 7 core parameters illustrated in Fig. 2 were calculated. F0 discontinuity is measured 1) between seg_1 and seg_2, which reflects the pitch reset properties of prosodic boundaries and for feature *f* it is denoted f_{1_2}, and 2) between each of these segments and the joint segment $seg_{12,}$ features denoted f_{1_12} and f_{2_12}, capturing the deviation of the pre- and post boundary F0 from a common tendency. Features with 'd' refer to difference between endpoints of the lines and those with 's' to difference between slopes. Hence, for example, d_{1_2} is the absolute F0 distance between the end point of the regression line in seg_1 and the start point of the corresponding line in seg_2; s_{1_12} is the absolute slope difference between the regression lines in seg_1 and the joint

seg_{12}. Using the midline and range fits as inputs the 7 parameters yield 14 features plus two for median F0 from the 500ms of the midlines before and after the boundary.

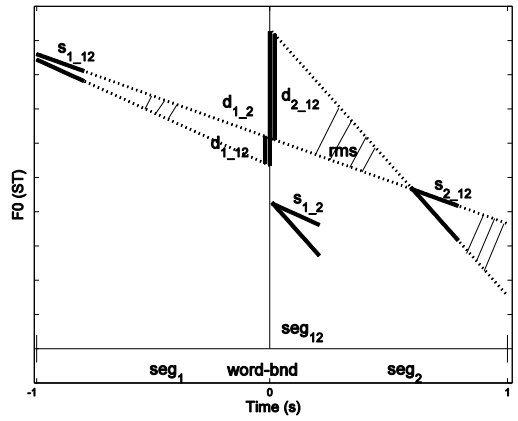

Figure 2: *Discontinuity parameters from stylization.*

For weak prosodic boundaries, 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*).

3. Results

3.1. Local and global resets in Lombard speech

We first look at the relationship between the continuous proxy of the Lombard effect in our data (*dct1*) and the reset features described above. Table 2 lists means of adjusted R^2 values for separate subjects from the regression models with *dct1* as an independent variable, the stylized features as the dependent variables, and boundary type as predictor.

Table 2. *Mean adjusted R^2 values from regression models, see text for details.*

input	d_{1_2}	d_{1_12}	d_{2_12}	s_{1_2}	s_{1_12}	s_{2_12}	*rms*
midline	0.63	0.4	0.48	0.51	0.38	0.4	0.44
range	0.18	0.33	0.22	0.13	0.14	0.16	0.4

We see that midline features provide consistently higher values than the range features, and thus offer a better feature set for capturing the relationship between raising pitch due to ambient noise and pitch reset in the three boundary types. In other words, resets in the pitch range across a boundary correlate only weakly with increasing F0 due to ambient noise.

To compare these global pitch reset values with a more local measure, we calculated the absolute value of the difference between F0 means of the hand-corrected pre-boundary nasal and post-boundary vowel (*resetN#V*) and from pre-boundary rhyme and first two post-boundary vowels (*resetR#V1V2*). The same models using these local pitch resets yield mean adjusted R^2 of 0.64 and 0.66 respectively. Hence, local resets, requiring hand correction, offer only slightly better characterization of noise-induced F0 scaling than global reset feature d_{1_2} (R^2 of 0.63) with no need for manual data-processing.

Fig. 3 shows the relationship between *dct1* on the one hand and midline d_{1_2} as the best global reset feature and *resetR#V1V2* as the best local feature on the other hand. Several observations can be made. First, both reset features provide good separation between B2 (falls) and the other two boundaries. However, the separation between B1 and B0 is weak, suggesting little difference in prosodic strength between a major (rising) boundary B1 and weak B0. Second, regarding the effect of Lombard noise, all lines with significant slopes in the right column have a positive relationship: increasing hyper-articulation increases local pitch reset. For subjects S1-S2, B1 rises tend to show the strongest effect while for S3-S5 it is the B2 falls. The situation with the global d_{1_2} feature is more complex and the slopes do not have a pre-dominant direction. Finally, comparing midline d_{1_2} with *resetR#V1V2*, some speakers get identical relationship with *dct1* (e.g. S4) whereas others show qualitatively different patterns (e.g. S2, and partly S3 with relatively high R^2 values in both plots).

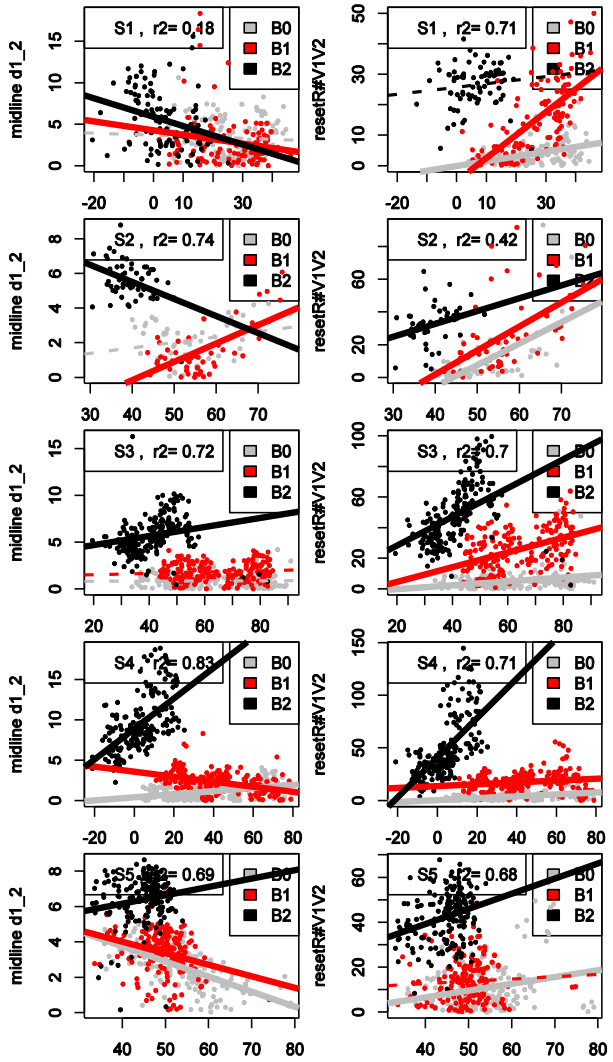

Figure 3: *Linear regression separately for subjects and three boundary types; solid bold lines have slopes significant at $p < 0.05$.*

Note that in Fig. 3, the regression lines in most cases (with an exception of speaker S2) get progressively more separated as *dct1* increases, i.e., with rising noise level. This is true in particular for separation between B2 and the remaining two boundary types. Therefore, it is possible, that reset features

will act as better predictors of boundary type in Lombard speech recordings compared to those obtained in silence.

3.2. Classification

For boundary level prediction we employed support vector machines (SVM) [2] with a linear Kernel function. The separating hyperplane was derived by sequential minimal optimization and accuracy is taken as mean performance of SVMs after tenfold cross-validation on held-out data with all 16 features from stylization described in Section 2.2. Table 3 presents accuracy of SVMs in classifying the boundary type for the five levels of ambient noise in the leftmost column and the confusion matrices from the classification.

The accuracy results suggest that ambient noise leading to hyper-articulation of F0 resets at prosodic boundaries is indeed beneficial for classifying the type of prosodic boundary with the stylization features, but that extreme hyper-articulation of 80nn condition does not have this effect.

Table3. *Confusion matrices from SVM predictions.*

Noise-cond		SVM prediction		
		B0	B1	B2
0	B0	0.756	0.156	0.089
	B1	0.405	0.481	0.115
69.028	B2	0.073	0.058	0.869
60	B0	0.804	0.104	0.092
	B1	0.242	0.630	0.127
74.624	B2	0.031	0.081	0.888
70	B0	0.740	0.212	0.048
	B1	0.277	0.622	0.101
72.796	B2	0.074	0.034	0.892
80	B0	0.841	0.123	0.037
	B1	0.474	0.474	0.053
75.002	B2	0.033	0.007	0.961
80nn	B0	0.559	0.386	0.055
	B1	0.306	0.571	0.122
69.709	B2	0.091	0.035	0.874

The matrix shows the best classification for B2 irrespective of the noise level and very little confusion with the other two types. On the contrary, B1 boundaries are classified the worst and commonly confused with B0, especially for "0" and "80" conditions. Lower accuracy for 0 and 80nn can be attributed to B1 mis-classified as B0 in the former and B0 mis-classified as B1 in the latter. These results corroborate the results from the previous section in the similarity of B0 and B1 in their response to ambient noise.

Although we focus here on F0 discontinuities, we ran a classification experiment including the pause duration between seg_1 and seg_2 among the feature set. While the accuracy increased significantly in each condition (peaking in 60 at 82.3) improving the classification of B0 the most, the miss-classification of B1 as B0 was not significantly improved.

Finally, we tested the hypothesis that with increasing noise and subsequent speaking up, F0 range features become 'saturated' in the sense that speaking up actually limits the range variability, which might explain the overall worse performance of range features compared to the midline level ones in previous section. We can infer the contribution of a feature toward boundary type identification with Silhouette measure defined for each data point i in (3) where $d_{A(i)}$ is the mean squared Euclidean distance of point i to all points of the same cluster and $d_{B(i)}$ is mean distance of point i to all points of the most i-similar cluster B not equal A.

$$(d_B(i) - d_A(i)) / \max(d_B(i), d_A(i)) \quad (3)$$

For us, the clusters are given by the boundary types, and we calculate for each instance i of the discontinuity feature in question, how well this instance is assigned to its related boundary level. The silhouette measure, however, does not support the hypothesis and there is no trade-off for assumed worse-performing range and better-performing midline level features with increasing noise.

4. Discussion & Conclusion

We have analyzed F0 discontinuity as a marker of prosodic boundary strength in three boundary types (weak B0, strong-rise B1, strong-fall B2) under increasing levels of ambient babble noise inducing hyper-articulation. The study offers several findings. First, reset features, both local and global, respond strongly to global F0 rising while this rise has smaller impact on weak boundaries. Boundaries B2 were minimally confused with other two boundary types irrespective of the noise level, but B1 and B0 showed considerable overlap. Also, within subjects, the response of reset to noise tended to be different for B2 and B1 with the latter being more similar to B0. This similar response of rises and weak boundaries to increasing noise might be due to 1) a salient difference in boundary strength between falls and rises, which might not be compatible with theoretical models suggesting identical strength for them, or, 2) the compensation of other markers of boundary strength for the differences in pitch discontinuities.

Second, it is not necessarily the case that increased ambient noise increases pitch reset of the boundaries; here subjects tend to employ rather varied strategies in how hyper-articulation affects the degree of discontinuity at the boundary.

Third, presence of noise improves the classification of boundary types but additional hyper-articulation induced by "artificial" means alleviates this effect. This might be related to a non-linear change in signaling boundaries for this extreme condition, and differences between strategies used by different speakers for the task to speak to a non-native speaker.

Fourth, concerning the applicability of the F0 discontinuities approach to Slovak Lombard speech shows that the feature set extracted from midline stylization performs better than features capturing the range discontinuities in boundary type classification. Also, the model might benefit from including the local F0 reset extracted from vowel intervals in the vicinity of the boundary if these are available.

Finally, although the beneficial effect of hyper-articulated Lombard speech on boundary classification is not entirely surprising, it is not a trivial consequence of F0 scaling, as the reset measures were computed in semi-tone scale. Rather, this phenomenon suggests that speakers do indeed amplify communicatively relevant features, in this case boundary reset characteristics, in presence of ambient noise, and that these adjustments are made in a predictable way.

5. Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055, and was also supported in part by VEGA grant 2/0197/15.

6. References

[1] Š. Beňuš and J. Šimko, "Stability and variability in Slovak prosodic boundaries," *Phonetica,* under review.

[2] C. Cortes and V.N. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, 1995.

[3] J. de Pijper and A.Sandermann, "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues," *J. Acoust. Soc. Am.*, vol. 96, pp. 2037–2047, 1994.

[4] D. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," Proc. 17the ICPhS, pp. 1233-1236, 2007.

[5] S.-A. Jun and J. Fletcher, "Methodology of studying intonation: From data collection to data analysis," in *Prosodic Typology II: The phonology of intonation and phrasing*, S.-A. Jun, Ed. Oxford: Oxford University Press, 2014, pp. 520–539.

[6] 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.

[7] M. Nespor and I. Vogel, *Prosodic Phonology*. Foris, Dordrecht.

[8] U. D. Reichel and K. Mády, "Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian," in *Proc. Interspeech 2014*, pp. 111—115.

[9] A. Savitsky and M.J.E.Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627—1639, 1964.

[10] E. O. Selkirk, "On derived domains in sentence phonology," *Phonology Yearbook*, vol. 3, pp. 371-405, 1986.

[11] J. Šimko, M. Vainio and Š. Beňuš, "Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue," *J. Acoust. Soc. Am.,* under review.

[12] M. Swerts, "Prosodic features at discourse boundaries of different strength," *J. Acoust. Soc. Am.,* vol. 101, pp. 514–621, 1997.

[13] M. Wagner and D. Watson, "Experimental and theoretical advances in prosody: A review," *Language and Cognitive Processes,* vol. 25, pp. 905–945, 2010.

[14] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, P. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Am.,* vol. 91, pp. 1707–1717, 1992.