Kinematic signatures of prosody in Lombard speech

þÿ B e H u a , ` t e f a n

International Speech Communications Association
2017

þÿ B e H u a , ` , S i m k o , J & L e h t i n e n , M 2 0 1 7 , K i n e m a t i c s i g n a t u r e s o f p r
speech . in Proceedings of Interspeech 2017 . Proceedings of the Annual Conference of the
International Speech Communication Association , International Speech Communications
Association , pp. 3013-3017 , Interspeech 2017 , Stockholm , Sweden , 20/08/2017 . https://doi.org/10.21437/Intersp

http://hdl.handle.net/10138/233902
https://doi.org/10.21437/Interspeech.2017-722

publishedVersion

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Kinematic signatures of prosody in Lombard speech

Štefan Beňuš[1], *Juraj Šimko*[2], *Mona Lehtinen*[2]

[1]Constantine the Philosopher University in Nitra, Slovakia, II SAS, Bratislava, Slovakia
[2]University of Helsinki, Finland

sbenus@ukf.sk, juraj.simko@helsinki.fi, mona.lehtinen@helsinki.fi

Abstract

Human spoken interactions are embodied and situated. Better understanding of the restrictions and affordances this embodiment and situational awareness has on human speech informs the quest for more natural models of human-machine spoken interactions. Here we examine the articulatory realization of communicative meanings expressed through f0 falling and rising prosodic boundaries in quiet and noisy conditions. Our data show that 1) the effect of environmental noise is more robustly present in the post-boundary than the pre-boundary movements, 2) f0 falls and rises are only weakly differentiated in supra-laryngeal articulation and differ minimally in their response to noise, 3) individual speakers find different solutions for achieving the communicative goals, and 4) lip movements are affected by noise and boundary type more than the tongue movements.

Index Terms: prosodic boundaries, Lombard speech, Slovak

1. Introduction

Understanding the encoding and decoding of communicative intensions through the variation in prosody is an ongoing long term goal of speech research. The present paper contributes to this overall goal by examining the articulatory correlates of prosodic boundary types when speech is produced in ambient noise. We are interested in the interplay between the linguistic and environmental requirements on the realization of prosody and in particular in the individual strategies how to find solutions to competing demands and achieve communicative goals. We thus wish to contribute to better understanding of situational awareness of humans regarding ambient noise and communicative intentions so that human-like qualities of automatic dialogue systems might be further improved.

Articulatory kinematics help us understand the effects of linguistic structure on the phonetic realizations; e.g. [1][2]. In lip kinematics, [3] found that structurally different prosodic positions are realized with different kinematic patterns, and that modelling this variation dynamically requires the interplay of multiple parameters. [4] explored how the degree of disjuncture of strong prosodic boundaries (IPs) is realized by the tongue's articulatory movements in the vicinity of these boundaries and subsequently examined the perceptual relevance of such articulatory signatures of prosodic strength. The variability of IP boundary strength was not syntactically motivated. They observed fine-grained, systematic, but individually different, strategies for communicating the boundary strength in the pre-, and especially cross-, boundary material, and corresponding perception of boundary strength expressed with these cues.

In a series of studies, [5],[6] explored the effects of babble noise on the prosodic realization in Slovak. Analyzing *global* articulatory adjustment of speakers to increasing levels of noise, [6] found support for the hypothesis that prosodically-driven hyper-articulation in noisy conditions is primarily effected by the less physiologically constrained lip-jaw system while the tongue provides the compensation for the increased f0 and lip-jaw movements stabilizing thus segmental contrasts in hyper-articulated speech. Analyzing the *local* durational and f0 correlates of the falling and rising prosodic boundaries and their hyper-articulation due to ambient noise, [5] found that 1) the realization of the falls and rises in quiet condition provides affordances for the patterns of strengthening in noise, and 2) falls were more strengthened than rises in f0 measures (height, range, cross-boundary reset) and pre-boundary lengthening, while cross-boundary pause length showed a weakening effect. Interestingly, the only articulatory measure considered in that study, jaw displacement, did not show any consistent effect.

Given the review above and advances in understanding the effect of Lombard speech on intelligibility [7-9], our goal in this paper is two-fold. First, we are interested in how individual speakers cope with the complex demands of communicating in noise and at the same time differentiating the falls from rises. Do they behave uniformly or do they find different solutions for this complex requirements? Second, the effects of ambient noise on prosodic changes associated with specific communicative needs such as signaling information structure have been investigated only in a limited scope [10]. We thus expand existing understanding of the production of prosodic variation by including both the lip and the tongue kinematics, comparison of boundaries with similar strength but clearly contrasting pragmatic meanings (falls vs. rises), and offer data of a less researched language (Slovak) that complement previous studies of global articulatory and local acoustic signatures of prosodic structure contributing to the picture of the cognitive system underlying the prosodic variation.

2. Methods

Data for this study come from the audio and electromagnetic (EMA) recordings of 5 Slovak speakers (3F, 2M) described in full in [5],[6]. Briefly, subjects read 12 balanced sentences designed to elicit one weak and two strong (Fall, Rise) boundaries in $V_0C_1\#V_1bV_2$ ($V_0 \in$ {[a:], [i:]}, $C_1 \in$ {[m], [n]}, $V_1 \in$ {[a], [i]}, $V_2 \in$ {[i], [a]}) such that the V-to-V transitions always included different vowels. In this paper we analyze the subset of that data where C_1 = [m] and the boundary is Fall or Rise: sequences […a:m#iba…] and […i:m#abi…]. In both sequences the pre-boundary rhyme is lexically stressed, realized with a pitch accent, and the segmental material in the vicinity of the boundary is identical for the Rises and Falls.

Subjects produced sentences in randomized blocks differentiated by the volume level of multi-speaker babble noise (mean f0 175Hz) administered over the headphones. In the reference block (quiet) subjects spoke naturally without any

http://dx.doi.org/10.21437/Interspeech.2017-722

noise or headphones. Three dB(A) levels of noise were used – 60, 70, and 80 dB – plus two other conditions eliciting extreme hypo- and hyper-articulation. Only quiet and 80dB conditions are analyzed in this paper.

Here we focus on the articulatory realization of the boundaries observable in 2D kinematic trajectories of sensors attached to the active articulators (lips, jaw, tongue) extracted after correcting for head movements. Three movements are used here: pre-boundary m-closing (*Mcl*, i:m#, a;m#), and two post-boundary ones: b-closing (*Bcl*, #ib(a), #ab(i)) and V_1-V_2-transition (*V2V*, #i(b)a, #a(b)i). *Mcl* and *Bcl* are assessed from the Lip Aperture (LA) variable between the two lip sensors and are defined by the kinematic landmarks of appropriate local minima in LA velocity profiles. The vocalic transition movement is defined similarly and based on the first principal component of the movement of the sensor attached to the tongue body. For each of the three movements, we consider its displacement (*Disp*), duration (*Dur*), and peak velocity (*PV*).

These kinematic variables are calculated after applying the normalization procedure for each movement described in [6]. *hh-index* is the trajectory of each movement in noise conditions relative to the average trajectory length for the same articulator for the same sentence uttered in the quiet condition. Hence, *hh-index* captures the situated relative responsiveness of individual articulators to given external conditions.

The primary independent variables of interest thus include two levels of noise (quiet vs. 80dB) and two types of strong prosodic boundary (Rise vs. Fall) in 558 tokens (2 sequences, 2 boundaries, 2 noise conditions, 10 or 20 repetitions per subject).

3. Results

We present first the overall patterns in the effect of Lombard noise and prosodic boundary on articulatory variables (mixed-effects modelling, individual subjects as random factors). Then we analyze the data in a speaker-dependent fashion to gauge the robustness of the patterns and potential individual solutions for the requirements imposed in the experimental procedure.

3.1. General patterns

The results of linear mixed effects modelling for the effects of noise (2nd column), boundary (3rd column) and their interaction (4th column) in the 9 dependent variables are summarized in Table 1. The data show a significant effect of noise on the displacement and velocity of the movements in the post-boundary domain (Bcl, V2V). These movements are more displaced and faster in noise than in quiet. The normalizing procedure in creating *hh-index* derived measures allows for assessing the sizes of the effect; e.g. m-closing displacement is greater in 80dB noise than in quiet by 25% on average. The comparison of Bcl and V2V shows that lip aperture in b-closing expands and fastens by almost 50% whereas the tongue vocalic transitions are expanded and faster by 7% only.

Finally, the interdependencies among the kinematic measures can be observed. For example, people robustly increase lip aperture in noise for post-boundary vowels (47% Bcl-Disp) and since this movement is minimally lengthened (6% Bcl-Dur), this expansion is carried out primarily by increasing velocity (46% Bcl-PV). The pre-boundary m-closing movement is expanded and faster to a smaller extent than the post-boundary one, but duration is lengthened more.

The pre-boundary m-closing movement is on average slightly expanded, longer and faster, but this effect does not reach significance. Similarly, all three movement durations yield positive estimates, being longer in noise than in quiet, but not significantly so. Both observations might be tied to variation among speakers and will be taken up in Section 3.2.

Table 1: *Model estimates: effect of Noise (quiet vs.80), Boundary (Fall vs. Rise) and their interaction on dependent variables; '*' is significant at p < 0.05, '?' tendency at p < 0.1.*

DepVar hh-norm	Noise (80)		Boundary (Fall)		Noise x Boundary	
Mcl-Disp	0,25	2,71?	-0,13	-1,82		
Bcl-Disp	0,47	4,53*	0,23	4,73*		
V2V-Disp	0,07	3,68*	0,07	1,56		
Mcl-Dur	0,12	2,38	0,06	1,08		
Bcl-Dur	0,06	1,93	-0,01	-0,46		
V2V-Dur	0,03	0,96	0,04	1,11		
Mcl-PV	0,16	1,18	-0,14	-2,28		
Bcl-PV	0,46	4,74*	0,16	6,17*		
V2V-PV	0,07	3,21*	0,08	2,16	-0,06	-3,19*

The pattern in differentiating Rises and Falls is more complex. Note that we expect either a similar effect of noise on them (both are IP boundaries with similar strengths), or the pattern showing slight strengthening of Falls compared to Rises (reflecting the f0 strengthening for Falls observed in [5]). First, if we consider the pattern in 80dB as 'strengthening', then Falls, compared to Rises, are stronger boundaries. The post-boundary b-closing is significantly more expanded and faster, and vocalic trajectory is also expanded and faster by similar increase when compared to the difference between quiet and 80dB noise (7% and 8% respectively), but this did not reach significance.

Second, Falls are slightly lengthened compared to Rises, but in pre-boundary domain Falls tend to be less displaced, and slower than Rises (negative, albeit non-significant values for Mcl-Disp and Mcl-PV) whereas Falls in the post-boundary movements (Bcl, V2V) are more displaced and faster. Hence, the communicative value associated primarily with f0 contours (Falls vs. Rise) is realized articulatorily asymmetrically regarding the pre- and post-boundary strengthening.

Finally, the interaction between the two factors obtain only in a single case (V2V-PV). Hence, the effects of increased noise and boundary are additive for most of the variables and there is minimal difference in how Falls and Rises respond to noise.

3.2. Individual patterns

Several patterns in the overall model did not reach significance and we suspected high individual variation among the speakers. We now look at individual speakers' resolving the requirements to convey boundary type in noise. Figure 1 shows the data and the caption explains the illustrations.

3.2.1. Response to noise

In the left panels, a group of speakers S4, S5 and S6 behaves very similarly in adjusting their lip articulation to ambient noise increase in both pre- and post-boundary consonants. They all expand the movement trajectory (S5 to a significantly greater degree than S4 and S6), but only minimally increase the duration (although this relative expansion is significant for S5 in both consonants and for S4 post-boundary). Although qualitatively similar, the relative expansion of kinematic measures, in particular for displacement and peak velocity is

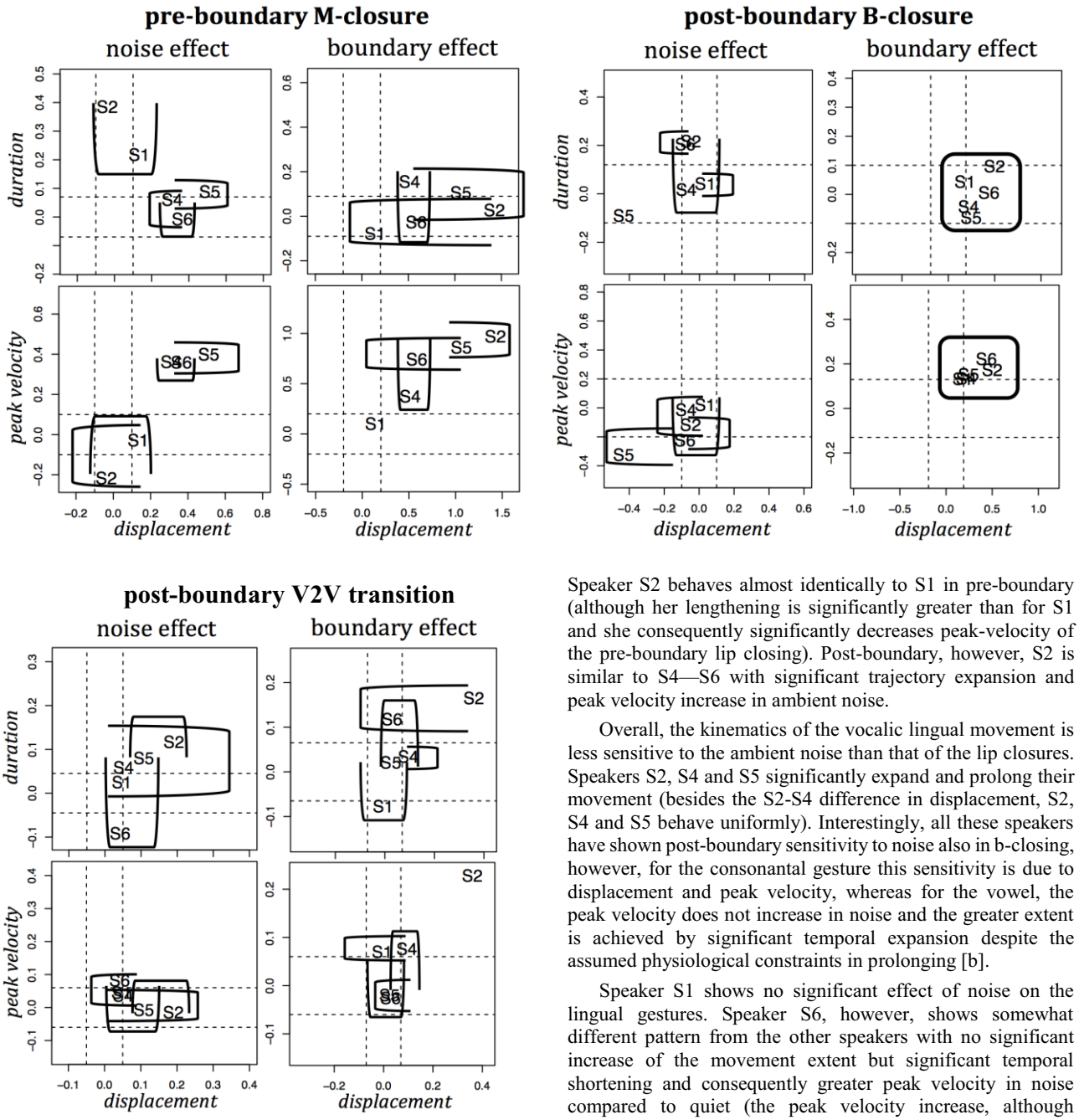

Figure 1: *Estimates of individual subjects' response of displacement, duration, and peak velocity to increased noise (left panels) and Fall (vs. Rise, right panels) in m-closing (top left), b-closing (top-right) and the vocalic transition (bottom). Dotted lines illustrate the intervals of non-significant variation and U-shaped symbols group subjects with non-significant differences among themselves; vertical for the x-axis and horizontal for the y-axis.*

greater post- than pre-boundary as can be seen by comparing the scales on the x-axes.

Speaker S1, on the other hand, shows no effect for post-boundary condition and minimal, albeit significant, relative increase in movement duration and displacement pre-boundary. Speaker S2 behaves almost identically to S1 in pre-boundary (although her lengthening is significantly greater than for S1 and she consequently significantly decreases peak-velocity of the pre-boundary lip closing). Post-boundary, however, S2 is similar to S4—S6 with significant trajectory expansion and peak velocity increase in ambient noise.

Overall, the kinematics of the vocalic lingual movement is less sensitive to the ambient noise than that of the lip closures. Speakers S2, S4 and S5 significantly expand and prolong their movement (besides the S2-S4 difference in displacement, S2, S4 and S5 behave uniformly). Interestingly, all these speakers have shown post-boundary sensitivity to noise also in b-closing, however, for the consonantal gesture this sensitivity is due to displacement and peak velocity, whereas for the vowel, the peak velocity does not increase in noise and the greater extent is achieved by significant temporal expansion despite the assumed physiological constraints in prolonging [b].

Speaker S1 shows no significant effect of noise on the lingual gestures. Speaker S6, however, shows somewhat different pattern from the other speakers with no significant increase of the movement extent but significant temporal shortening and consequently greater peak velocity in noise compared to quiet (the peak velocity increase, although significant, is not significantly different from S1 and S4).

3.2.2. Boundary patterns

As suggested by the mixed effect models, all speakers very similarly differentiate the boundary type by articulatory strengthening expanding trajectory and increasing peak velocity – for post-boundary lip closure in Falls compared to Rises. On the other hand, duration of this gesture is not significantly different between Falls and Rises. In fact, there are no significant differences for the relative change in the kinematic characteristics between the speakers for this gesture (although the relative increase in trajectory for Falls is not significant for S1 and the peak velocity increase for S1 and S2).

In contrast, pre-boundary, the consonantal closure in the stronger Fall boundary does not have relatively greater

duration; for S5, the gesture is actually significantly spatially smaller in Falls than Rises; for all other speakers there is no significant difference in gesture duration. For S5 this pre-boundary weakening of the stronger boundary is accompanied by significant relative decrease in peak velocity but not duration. Speakers S2 and S6, on the other hand, have significantly longer gestures in Falls compared to Rises, speaker S6 consequently significantly decreases its peak velocity in Falls (for S2 this decrease fails to reach significance, although it is not significantly different from S6). S1 and S4 do not significantly differentiate the pre-boundary gesture in any kinematic characteristic between the two boundaries.

The behavior of speakers in differentiating the boundary type by characteristics of the vocalic gesture is particularly inconsistent. First of all, S2 differs from all speakers both qualitatively and quantitatively by the rather substantially greater trajectory, duration and peak velocity in Falls than in Rises (although the difference in durational expansion is not significantly different from that of S6). For S5, there is no significant difference between the boundaries in the lingual gesture. Speaker S6 uses significantly temporally longer, but neither faster nor spatially greater, gesture in Falls. For S1, the tongue movement is significantly shorter and faster in Falls than in Rises. Finally, for S4, the gesture in Falls is significantly spatially greater that in Rises (although this increase is not significantly different from S5 and S6) and faster (again, however, not significantly more so that for S5).

3.2.3. Interactions

There are 7 significant interactions (out of 45 models overall) between the noise and boundary effects, all for speakers S2 and S5 with the quantitatively strongest effect of noise. All interaction effects are negative; therefore the noise effect in Falls is relatively smaller than additive effect of both noise and boundary. When the main boundary effect is positive this means that noise simply lessens the strength of the boundary effect in the loud condition; this is the case for S5's vowel displacement and S2's b-closing displacement and peak velocity. For vowel peak velocity measure (for both S2 and S5; this is the only case when the interaction is significant in the mixed effect models) the interaction in fact reverses the sign of the boundary effect: in quiet the peak velocity is greater in Falls than Rises, in 80 dB it is the other way around. For the m-closing displacement and peak velocity for S5 the boundary main effect is negative, meaning that the relative decrease in these measures in Falls is further magnified in 80 dB noise.

Finally, counting the significant post-hoc differences, speakers are less consistent in their response to noise (41 out of 120 between-subjects differences are non-significant) than to Boundary (69/120). This, appended with few interactions, suggests that the linguistic communicative requirement (Fall vs. Rise) is produced more consistently than the para-linguistic response to noise. This is despite the fact that the overall response to noise tends to be more uniform (all positive values in Table 1) than expressing continuation and finality.

4. Discussion & Conclusions

We set out to explore the response of the articulatory system to the communicative requirements to convey finality or continuation (Fall vs. Rise) in quiet and noisy environments. Our results show several patterns. First, lip movements are affected by noise and boundary type more than the tongue movements. This may stem from the distance of the movement from the boundary in our stimuli (B-cl being closer than V2V) and is in line with a positive relationship between the effect of boundary strength on movement's realization and the distance of the movement from the boundary [11]. However, it might also be due to the greater restrictions of the tongue movement compared to the lips for expressing prosodic meanings [6].

Second, systematic general patterns arise from the consistency of individual speakers, and the absence of such patterns is due to either the lack of the effect in the speaker's behavior or the systematic differences among the several possibilities available to the subjects for resolving the communicative requirements. We found evidence for all three situations in our data. Cuing the boundary type in post-boundary b-closing movement illustrates the general systematic pattern followed by all speakers. Pre-boundary m-closing and its response to noise shows the pattern where subjects follow consistent but significantly different strategies, which results in the overall absence of the significant pattern across subjects.

Third, the most consistent strengthening for both noise and falling boundary type are shown in the post-boundary material, and, with the displacement and velocity rather than duration of the movement. This contrasts somewhat with [4] who found cross-boundary movements the most robust. (Labeling our cross-boundary m-opening was less reliable than the other three movements and is thus not included in the results, but available data show a similar effect sizes as post-boundary b-closing.)

Note that pre-boundary positions in our stimuli are pitch accented, and crucial for conveying the Fall/Rise meaning, whereas post-boundary ones are commonly prosodically non-prominent. Hence, the lack of strong linguistic constraints in the post-boundary material might facilitate the greater affordance for its response to situated communicative needs. Relatedly, noise affects the articulatory signatures of prosodic boundaries more than the boundary type and thus the boundaries are realized more consistently. This again might be construed as evidence for the greater constraints of the linguistic requirements (Fall/Rise) than environment (noise) on the variability in the prosodic system.

Also, [3] observed that accenting results in movements that are "simply bigger in all ways – in distance, time, and speed whereas boundary-induced strengthening effects are evident in longer, but not necessarily faster, articulation in both domain-initial and domain-final positions." In our data, noise-induced strengthening in the vicinity of boundaries tends to be 'bigger in all ways' but with a systematic pattern (greatest spatial expansion, smaller durational lengthening, and medium velocity increase) and consistent subject differences. It seems that noise-induced strengthening of prosodic boundaries is neither like accent- nor like boundary- strengthening.

Fourth, corroborating [4]'s observation of systematic difference between IPs of different communicative intentions, we found consistent differences between Falls and Rises of similar boundary strength. We extend their results by observing the lack of interactions between noise and boundary type and thus minimal differences in how ambient noise affected the Falls/Rises contrast both overall and in subject-wise fashion.

5. Acknowledgements

This work has been supported by Scientific Granting Agency in Slovakia (grant 2/0197/15) and by the Academy of Finland Digital Language Typology project (No. 129).

6. References

[1] D. Byrd, A. Kaun, S. Narayanan and E. Saltzman, "Phrasal signatures in articulation", In: M.B. Broe and J.B. Pierrehumbert (Eds.), *Papers in laboratory phonology 5: Acquisition and the lexicon*, pp. 70–87.Cambridge University Press, 2000.

[2] J. Edwards, M. Beckman and J. Fletcher, "The articulatory kinematics of final lengthening", *Journal of the Acoustical Society of America* 89, pp. 369–382, 1991.

[3] T. Cho, "Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English," In L. Goldstein, D. H. Whalen, C. T. Best (Eds.), Laboratory Phonology 8, pp.519-548. Mouton de Gruyter, 2006.

[4] J. Krivokapić and D. Byrd, "Prosodic boundary strength: An articulatory and perceptual study", *Journal of Phonetics* 40, pp. 430-442, 2012.

[5] Š. Beňuš and J. Šimko, "Stability and variability in Slovak prosodic boundaries," *Phonetica* 73(3-4), 163-193, 2016.

[6] J. Šimko, Š., Beňuš and M. Vainio, "Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue," *Journal of the Acoustical Society of America*, 130(4), 2116-2127, 2016.

[7] Y. Lu and M. Cooke, M. "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, 51(12), 1253-1262, 2009.

[8] M. D. Skowronski a J. G. Harris."Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, 48(5), 549-558, 2006.

[9] M. Cooke, C. Mayo a J. Villegas, J. "The contribution of durational and spectral changes to the Lombard speech intelligibility benefit," *The Journal of the Acoustical Society of America*, 135(2), 874-883, 2014.

[10] Arciuli, J., Simpson, B. S., Vogel, A. P., & Ballard, K. J. (2014). Acoustic changes in the production of lexical stress during Lombard speech. Language and speech, 57(2), 149-162

[11] D. Byrd, J. Krivokapić and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effects", *Journal of the Acoustical Society of America* 120, pp. 1589-1599, 2006.