Interspeech 2018
2-6 September 2018, Hyderabad

Articulatory consequences of vocal effort elicitation method

Elísabet Eir Cortes[1], *Marcin Włodarczak*[1], *Juraj Šimko*[2]

[1]Department of Linguistics, Stockholm University, Stockholm, Sweden
[2]Department of Digital Humanities, University of Helsinki, Helsinki, Finland
{elisabet.cortes | wlodarczak}@ling.su.se, juraj.simko@helsinki.fi

Abstract

Articulatory features from two datasets, Slovak and Swedish, were compared to see whether different methods of eliciting loud speech (ambient noise vs. visually presented loudness target) result in different articulatory behavior. The features studied were temporal and kinematic characteristics of lip separation within the closing and opening gestures of bilabial consonants, and of the tongue body movement from /i/ to /a/ through a bilabial consonant. The results indicate larger hyper-articulation in the speech elicited with visually presented target. While individual articulatory strategies are evident, the speaker groups agree on increasing the kinematic features consistently within each gesture in response to the increased vocal effort. Another concerted strategy is keeping the tongue response considerably smaller than that of the lips, presumably to preserve acoustic prerequisites necessary for the adequate vowel identity. While the method of visually presented loudness target elicits larger span of vocal effort, the two elicitation methods achieve comparable consistency per loudness conditions.

Index Terms: articulation, elicitation methods, Lombard speech, varying vocal effort, visually presented loudness target

1. Introduction

Varying vocal effort is part of a set of skills associated with spoken language production. It involves the speaker varying her or his subglottal pressure, as well as adjusting the laryngeal muscles to accommodate this increased pulmonic flow. Even though vocal effort can be varied from whisper to shouting, this paper only deals with vocal effort in phonated speech. Varying vocal effort has been studied both on its own and in connection with the accompanying changes in fundamental frequency and voice quality [1-6]. A particularly well studied phenomenon is the Lombard effect [7]. It is the increase in vocal effort that automatically follows if the speaker is subjected to noisy surroundings. Changes in pitch and some articulatory settings are also a part of this effect, see [8] for an overview. Loud speech produced without ambient noise has been somewhat less studied, see [3-6], and to the best of the authors' knowledge only two studies seem to have been published in which this style of increased effort has been compared to Lombard speech [9, 12].

Recent years' technological progress in methods for tracking articulatory movements has resulted in a new series of studies focusing on the kinematics of speech. One such series proposes Lombard speech as a reliable method to elicit even and predictable increments in loudness [10, 11]. An alternative method to elicit increased loudness is one where the speakers are not directly instructed to raise their voice but do so by reaching a visually presented loudness target. Different elicitation methods have been shown to evoke different behavior in the speaker, with respect to respiratory strategies [12] and possibly also in the acoustic output (f_0 and energy) [13].

The aim of this paper is to investigate possible differences in the articulatory characteristics of speech with varying vocal effort elicited by the two different methods described above: the Lombard effect on one hand, and self-monitored vocal effort with visual aid, on the other. The main questions addressed are whether articulation is sensitive to elicitation method, and in that case what kind of different articulatory behavior each elicitation method evokes. As well as adding to our understanding of speech motor control, answering these questions is important for studies on articulation, where it for instance might be crucial to discern between articulatory behaviour attributable to the elicitation method as opposed to the factors under observation.

2. Methods

2.1. Elicitation methods

Data for this study come from two articulatory datasets, one with Slovak speakers, and one with Swedish speakers. The Slovak data consists of articulation and acoustics of five native speakers reading real-word sentences in a fluent style. Recordings were done in Helsinki with the AG500 system (Carstens Medizinelektronik GmbH, Germany). This dataset has previously been described in full in [10] and [11]. The Swedish data set consists of audio and EMA (electromagnetic articulography) recordings performed in Stockholm (The Wave system, Northern Digital Instruments [NDI], Canada), as well as electroglottographic (EGG) and video recordings, of seven native speakers of Swedish. The speakers read lists of i.a. nonsense phrases figuring each of the Swedish vowels, stressed, in an identical context.

The difference between the two data sets of greatest importance for this paper, concerns the methods used to elicit varying vocal effort. For the Slovak set, multi-speaker babble noise of 60, 70, and 80 dB played through headphones was used to evoke Lombard speech. In addition, the speakers were recorded in a silent condition, producing a total of four loudness conditions. For the Swedish set, the speaker microphone output was routed through a dampening device located in an adjacent control room, and back into a red diode VU-meter, which served as a visual aid for the speakers to keep their vocal effort at a certain level. The speakers were instructed to make sure the red lights were visible and were informed that they might have to vary their vocal effort in order to achieve this. The visual aid was situated at the side of the presentation screen for the speech material so that the speakers could monitor the presence of red light from the corner of their eye. The speaker output was dampened in 10

10.21437/Interspeech.2018-1038

dB steps, each step requiring more vocal effort to keep the red lights visible. The dampening levels ranged from four to five, depending on each speaker´s effort capacity. This visually aided self-monitoring proved to be a setup in which the speakers did not have to invest mental energy into monitoring their loudness level but could instead focus on uttering the speech material.

The two data sets will in this paper be referred to according to their elicitation method, LOM (Lombard effect - Slovak data), and VIS (visual loudness target - Swedish data).

2.2. Material and analysis

Small audio and EMA subsets were chosen from the VIS and LOM datasets. The first subset consisted of bilabial-/a/-bilabial sequences (/bab/ in VIS, /mab/ in LOM), used for studying lip movements. The second subset consisted of /iba/ sequences, used for studying the movement of the tongue body from vowel to vowel. This resulted in the speech material presented in Fig. 1.

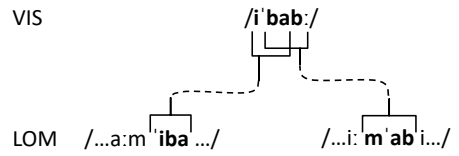

Figure 1: *The speech materials chosen from each of the datasets, and how they map to each other. Bold: selected speech sound sequences.* b: a: i: *are long sounds;* ˈ *indicates stress on the following syllable.*

In VIS, both the /bab/ and the /iba/ sequences were extracted from 99 repetitions of /iˈbab:/, with 11-18 repetitions per speaker. In LOM, the Swedish /bab/ sequences were matched with 296 repetitions of /mab/ extracted from /i:**m**ˈ**ab**i/ (17-79 repetitions per speaker). The Swedish /iba/ tokens were matched with 278 repetitions of /iba/ extracted from /a:mˈ**iba**/ (18-75 repetitions per speaker). All in all, the speech material consisted thus of a total of 198 VIS tokens across 4-5 loudness levels, and 574 LOM tokens across 3-4 loudness levels.

The movement of EMA sensors attached to the active articulators (upper lip, lower lip and tongue body) were extracted after correcting for head movements. Four movements are used here: closing and opening gestures of the first consonant (C1cl and C1op); closing gesture of the second consonant (C2cl); and the vowel to vowel transition (V2V). To assess C1cl, C1op and C2cl, a Lip Aperture variable (LA) was defined as the Euclidian distance between the two lip sensors. To assess the vocalic transition (V2V) the trajectory of the movement of the tongue body sensor was calculated. For each of the four movements, we quantify its kinematic characteristics by calculating its peak velocity (PV) as well as displacement (Disp) for the LA movements and trajectory length (Traj) for the V2V movement. The temporal characteristics of each movement is captured by its duration (Dur). The features were defined by the kinematic landmarks of appropriate local minima in the LA and the V2V velocity profiles. These articulatory variables then underwent a slightly adapted version of the hh-normalization procedure described in [10] and [11]. Specifically, the maximum intensity (dB SPL) of each speaker's output in the conversational (quiet/relaxed) condition was subtracted from the speaker's output in the other conditions, arriving at a relative intensity (Rel.Int) for each utterance. Subsequently, an hh-index was calculated for each of the features (Disp, Traj, PV, and Dur), in the following way: For each utterance and gesture the values were divided by the corresponding feature value for the same articulator in the speaker's utterance with Rel.Int closest to 0 dB. The outcome was a ratio for the tokens within each feature. This procedure captures the relative kinematic and temporal characteristics of individual articulators, thus normalizing differences between and within the two data sets arising from, for instance, the anatomy of the individual speaker, variation in sensor placement, etc.

3. Results

3.1. Exploratory data analysis

3.1.1. Intensity levels

Both methods elicit consistently increased intensity levels, see Fig. 2. While variation between the individual speakers can be seen, the elicitation methods achieve comparable consistency per loudness condition.

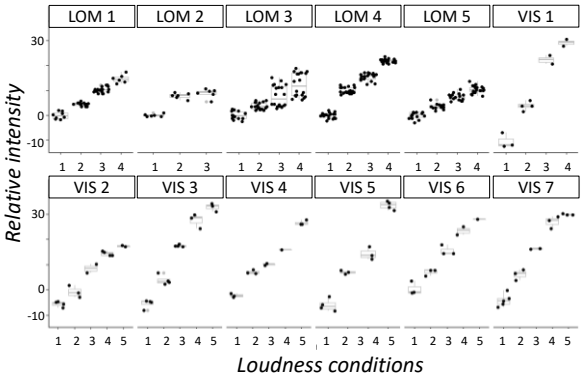

Figure 2: *Individual speakers' intensity levels. LOM 1-5: ambient noise, VIS 1-7: visual loudness target.*

The ranges of relative intensity achieved by the two different elicitation methods can be read from Fig. 2. Notably, the span resulting from the visual target method is greater in dB (M=35.76, SD=7.59, compared to M=18.75, SD=5.39 for the ambient noise method), reaching further towards both shouted and soft speech. As softer speech levels might be associated with articulatory behavior different from conversational levels [14], the levels below the reference levels of conversational vocal effort were excluded from further analyses.

3.1.2. Articulatory strategies

Individual articulatory strategies were observed, seemingly unrelated to elicitation method. Fig. 3 shows a clear example of inter-speaker variation in the peak velocity of LA (Lip Aperture) in the closing phase of the first consonant. LOM speakers 1 and 2, for instance, do not seem to exhibit the same increase in peak velocity as the remaining LOM speakers. VIS speaker 2 has a non-linear increase of PV, not as clearly observable in the other speakers.

Despite the inter-speaker variation observed in all examined features, some overall relationships could be discerned, as for instance the increase in PV with increased Rel.Int indicated by Fig. 3.

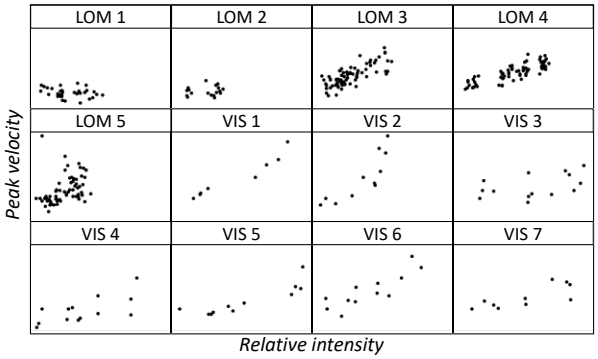

Figure 3: *Individual strategies for PV in LA over increasing Rel.Int in C1cl. LOM 1-5: Lombard speech. VIS 1-7: Visual target method.*

3.2. Statistical modelling

3.2.1. General effects

The relationships indicated by the exploratory data analysis were evaluated in R (v.3.4.2) [15], using *lmer* v. 2.0-33 [16] to perform a linear mixed effects analysis predicting hh-indexed features as a function of the fixed effects REL.INT, ELICIT (elicitation method), and REL.INT:ELICIT (their interaction). Each articulatory feature was modeled separately and the model releveled for ELICIT to get the effect on each speaker group. Given the presence of individual strategies, speaker slopes were entered as random effects into the models. Outliers in the distribution of residuals deviating by more than 2 standard deviations from their overall means were excluded before refitting the model. Table 1 summarizes the results.

Table 1: *Effect of relative intensity in Bel on the hh-normalized dependent variables, and its interaction with elicitation method.*

DEP.VAR hh-norm	REL.INT LOM		REL.INT VIS		REL.INT*ELICIT
C1clDisp	0.1948	*	0.2746	**	
C1clPV	0.2008	*	0.2983	**	
C1clDur	-0.0027		-0.0171		
C1opDisp	0.1838		0.4766	***	
C1opPV	0.1998		0.3763	**	
C1opDur	-0.0181		0.0656	*	*
C2clDisp	0.2199		0.4722	***	
C2clPV	0.2001		0.4269	***	
C2clDur	-0.0320		0.0237		*
V2VTraj	0.0314		0.1111	***	**
V2VPV	0.0322		0.1111	***	*
V2VDur	0.0065		0.0160		

The analysis revealed a significant effect of REL.INT on all kinematic features (Disp, PV, Traj), but in the VIS data (visual target) only. A significant REL.INT:ELICIT interaction was found only in the temporal feature (Dur) in C1op and C2cl, as well as in the kinematic features of V2V, indicating significant differences between speaker groups/elicitation methods with respect to the effect of REL.INT on these features. However, given that the slope estimates are expressed in Bel, the response in these features is small. Effect of REL.INT in the LOM data (Lombard speech) is only to be seen in the C1cl. More interestingly, the slope estimates for duration-related features are close to zero in almost all cases, pointing to a systematic articulatory response. Specifically, speakers seem to compensate for increased displacement by increasing peak velocity while keeping duration relatively constant. Moreover, the slope estimates for VIS are 2-3 times greater across features than for LOM (albeit not significantly different for all but four comparisons), which implies greater hyper-articulation in VIS.

3.2.2. Speaker-specific effects

A different approach was used to compare the effect of increased intensity on each articulatory feature on an individual speaker level. A linear regression analysis was performed to model the features as a function of REL.INT, individual speaker, and the interaction between these two variables. Each feature was modelled separately and releveled to individual speakers. Fig. 4 shows per-speaker slope estimates derived from these models. Each point in the plots denotes an individual speaker's rate of change in an articulatory feature in response to increased REL.INT, plotted against this speaker's rate of change in another articulatory feature. For instance, a speaker located high on any axis indicates that when she/he spoke louder, she/he exhibited high degree of change in the articulatory feature associated with that particular axis.

Four major observations can be made. Firstly, the rate of change in the kinematic features (Disp, PV) in all consonant gestures exhibit high correlation across the data: R^2 ranging from 0.86 to 0.95 (Fig. 4, 2nd row). The correlation of the corresponding features in the vocalic transition is a bit lower (R^2 = 0.61) which can be attributed to the weaker VIS correlation (R^2 = 0.17 vs. R^2 = 0.89 for LOM, see Fig. 4, last plot). Secondly, while the VIS speaker group has more consistently positive or negative slopes, the LOM group falls on either side of zero. This could explain the apparent lack of effect from increasing intensity in the LOM group in Table 1. Thirdly, the rate of change in Disp and PV in the vocalic transition is much lower than in the consonant gestures, on average 1:6 for the LOM data and 1:3 for the VIS data, see Fig. 4, 2nd row. Lastly, the rate of change for the temporal feature (Dur), exhibits less variance and weaker relationship with the kinematic features, see Fig. 4, 1st row.

4. Discussion and conclusions

There are some obvious differences between the data sets used in the present study. Besides the difference in elicitation methods, the two data sets differ in language (Slovak vs. Swedish), speech style (fluent speech vs. reciting), and utterance type (longer sentences vs. two-syllable nonsense words). Moreover, the speech material chosen for this study differs on phonetic as well as prosodic levels: The first consonant is the nasal /m/ in LOM, the plosive /b/ in VIS. The second consonant /b/ is long in VIS but not so in LOM. The prosodic differences are related to stress, which might have a systematic effect on the articulatory gestures. In VIS, the second syllable is stressed, while in LOM the first vowel in each sequence is stressed. Differences were also observed in the dynamic range employed, with larger span in VIS. The likely reason for this is that while response to noise results invariably in increased vocal effort, the method employed in VIS can be used to elicit both loud and soft speech. In

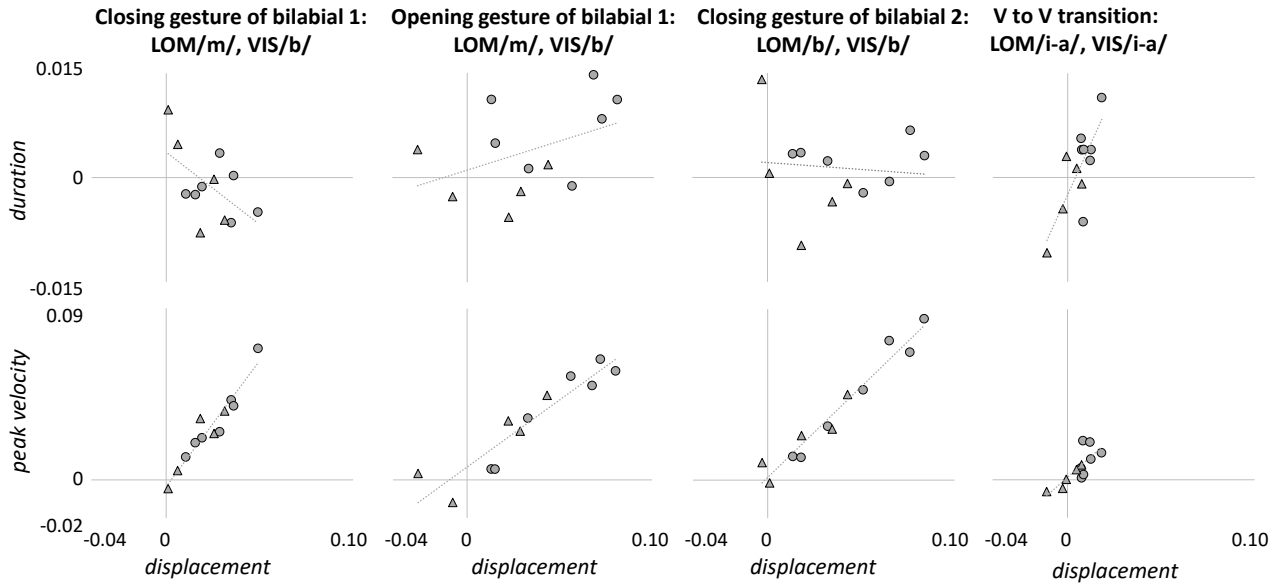

Figure 4: *Slope estimates (rate of change) of individual subjects' response of Disp, PV, and Dur (all hh-normalized) to* REL.INT *in C1cl (leftmost column), C1op (second column from left), C2cl (third column from left), and V2V (rightmost column). Triangles: LOM speakers (ambient noise), circles: VIS speakers (visual target). Dotted trend line: correlation for all speakers.*

addition, the highest intensity level of the Lombard speech was achieved by 80 dB noise presented through headphones; increasing this noise further might result in hearing damages.

Optimally, we would want to compare different elicitation strategies with the same speakers, language, speech material and EMA equipment; consequently, the results from the current study should be considered preliminary.

Yet, despite the differences at hand, similar articulatory strategies were observed across the datasets. Namely, speakers seem to aim to preserve the temporal structure of the whole utterance, which can be inferred from the lack of effect on duration. This comes at the expense of the peak velocity, whose rate of change is very well correlated with the rate of change in displacement/trajectory length (see Table 1 and Fig. 4). The preferred strategy thus appears to be to locally manipulate peak velocity, rather than stretch duration to accommodate the increased displacement accompanying increased vocal effort.

The slope estimates in Table 1 reveal a larger response in VIS, across almost all features. This suggests consistently stronger hyperarticulation in the VIS group compared to the LOM group, i.e. larger articulatory changes in response to increased intensity. The difference is also consistently greater for vowels than for consonants. Furthermore, the lack of significant effect of intensity in the LOM data (Table 1), might be explained by the behavior portrayed in Fig. 4, where the LOM speakers' slopes fall on either side of zero, pointing to more diverse individual articulatory strategies.

The comparatively low rates of change in the vocalic transitions in Table 1 and Fig. 4 are somewhat surprising. One would expect more displacement as the vowels grow louder, see for instance [4]. A likely explanation for this lack of effect is that the speakers strive to keep the effect of vocal effort on the tongue at a minimum with a view to maintaining the acoustic settings necessary for the intended vowel identities. A lowered tongue in /i/ would make it sound like an /e/ or even an /ɛ/. Presumably, while the increased vocal effort induces increased lip movements, the tongue compensates for the increased displacement. This could be investigated in some future studies, in which the vowels in the datasets are tested for identity, perceptually as well as subjected to formant analysis. The lip articulation in the vocalic transition would be taken into account, and the results from the lip and tongue displacements correlated to the outcome of the identity tests.

To summarize our conclusions, articulatory differences can be discerned between the data sets elicited by the two methods under comparison. Although significance was only present for the differences in some of the articulatory features, it can be inferred that the speech elicited with the visual target method displays consistently stronger hyperarticulation compared to the ambient noise method with respect to the durational and kinematic features studied here. Our observations also suggest that the magnitude of this difference across features is consistent within sound type, and greater in vowels compared to consonants. The results indicate that articulation might be sensitive to elicitation methods. Nevertheless, further studies on a more comparable material are still needed to shed more light on this issue.

Despite individual speaker strategies within both groups, some common articulatory strategies are evident. First, increased vocal effort results in increased displacement and peak velocity of the lip aperture, presumably serving the purpose of maintaining the temporal structure of the whole utterance. Second, tongue displacement and peak velocity are kept at a minimum in the vocalic transition, most likely to preserve adequate vowel identity in the face of hyper-articulation in the lips.

5. Acknowledgements

This work was in part supported by a Stockholm University grant for collaboration with University of Helsinki. We are very grateful to Štefan Beňuš and Mona Lehtinen for their crucial contribution to designing and recording the Slovak stimuli.

6. References

[1] Rostolland, D. “Acoustic features of shouted voice,” *Acustica*, vol. 50, pp. 118-125, 1982a.

[2] Rostolland, D. “Phonetic structure of shouted voice,” *Acustica*, vol. 51, pp. 80-89, 1982b.

[3] Traunmüller, H. and Eriksson, A. “Acoustic effects of variation in vocal effort by men, women, and children,” *Journal of the Acoustical Society of America*, vol. 107, pp. 3438-3451, 2000.

[4] Schulman, R. “Articulatory dynamics of loud and normal speech,” *Journal of the Acoustical Society of America*, vol. 85, pp. 295-312, 1989.

[5] Bonnot, J.-F. P. and Chevrie-Muller, C. “Some effects of shouted and whispered conditions on temporal organization,” *Journal of Phonetics*, vol. 19, pp. 473-483, 1991.

[6] Geumann, A. *Invariance and variability in articulation and acoustics of natural perturbed speech.* Ph.D. Dissertation. University of Munich, 2001.

[7] Lane, H. and Tranel, B. “The Lombard sign and the role of hearing in speech,” *Journal of Speech and Hearing Research*, vol. 14, pp. 677-709, 1971.

[8] Brumm, H. and Zollinger, S.A.”The evolution of the Lombard effect: 100 years of psychoacoustic research,” *Behaviour*, vol. 148, no. 11-13, pp. 1173-1198, 2011.

[9] Bond, Z., Moore, T.J., and Gable, B. “Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask,” *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 907–912, 1989.

[10] Beňuš, Š. and Šimko, J. “Stability and variability in Slovak prosodic boundaries,” *Phonetica* vol.73, no. 3-4, pp. 163-193, 2016.

[11] Šimko, J., Beňuš, Š., and Vainio, M. “Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue,” *Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2116-2127, 2016.

[12] Huber, J.E. “Effect of cues to increase sound pressure level on respiratory kinematic patterns during connected speech,” *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 3, pp. 621-634, 2007.

[13] Bond, Z.S. and Moore, T.J. "A note on loud and lombard speech," In *ICSLP-1990*, pp. 969-972., 1990.

[14] Davis, C. and Kim, J. “Whispered and Lombard speech: different ways to exaggerate articulation,” *SST2016*, 6–9 December, Parramatta, Australia, 2016.

[15] R Core Team, “R: A language and environment for statistical computing,” R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[16] Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. “lmerTest Package: Tests in Linear Mixed Effects Models,” *Journal of Statistical Software*, vol. 82, no. 13, pp. 1-26, 2017.