Gesture Duration and Articulator Velocity in Plosive-Vowel-Transitions

Dominik Bauer[1], Jim Kannampuzha[1], Phil Hoole[2], and Bernd J. Kröger[1]

[1] Department of Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Aachen and RWTH Aachen University, Aachen, Germany
[2] Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, Munich, Germany
dobauer@ukaachen.de, jkannampuzha@ukaachen.de, hoole@phonetik.uni-muenchen.de, bkroeger@ukaachen.de

Abstract. In this study the gesture duration and articulator velocity in consonant-vowel-transitions has been analysed using electromagnetic articulography (EMA). The receiver coils where placed on the tongue, lips and teeth. We found onset and offset durations which are statistically significant for a special articulator. The duration of the offset is affected by the degree of opening of the following vowel. The acquired data is intended to tune the control model of an articulatory speech synthesizer to improve the acoustic quality of plosive-vowel-transitions.

Keywords: Speech, articulatory speech synthesis, articulation, electromagnetic articulography, EMA.

1 Introduction

The articulators in a human vocal tract are physically affected by inertia, which causes a transitional phase between two successive articulatory constellations. All these movements are coded in the acoustical signal. In a digital model of the vocal tract, velocity and acceleration of articulators must be modelled somehow to provide natural transitions from one phone to another. One method to deal with this problem is to capture the articulator velocity and the gesture durations in different phone contexts from a human speaker. The gesture-based control of an articulatory speech synthesizer requires detailed knowledge of the synchronization and the speed of articulatory action units [1], [2]. In our recent work, we analysed the synchronization of action units based on acoustic observations [3]. Since some articulatory events like the begin of a consonantal movement (before the articulator really forms a constriction) are hard to estimate from an acoustic representation of the signal, we decided to perform an EMA-study to measure the articulator speed in consonantal stops.

The control model of Kröger and Birkholz divides the articulatory movement into an onset and offset interval [4], [2]. These intervals can also be found in the EMA representation of articulatory movements. (For a detailed description of electromagnetic articulography see [5].) In Figure 1, the onset and offset of an apico-alveolar

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNCS 5967, pp. 346–353, 2010.

stop movement in the utterance "ich habe [de:] gesagt" are highlighted. The upper two tiers show the oscillogram and spectrogram, the lower two trajectories show tongue tip position and tongue tip speed.

In the onset interval, the tongue tip height is strictly increasing (velocity > 0). After a short (quasi-) stationary phase, where the tongue is in contact with the palate (velocity = 0) the constriction is released. During this offset phase, the tongue height is strictly decreasing (velocity < 0). It is easy to see that the onset transition already starts when the preceding vowel becomes audible. Similarly, the offset reaches into the following vowel. In this study we measured the duration of these intervals in different vowel-contexts to obtain information on consonantal timing for use in articulatory speech synthesis control. From our results we expect an advantage in naturalness and intelligibility in plosive-vowel transitions in articulatory speech synthesis.

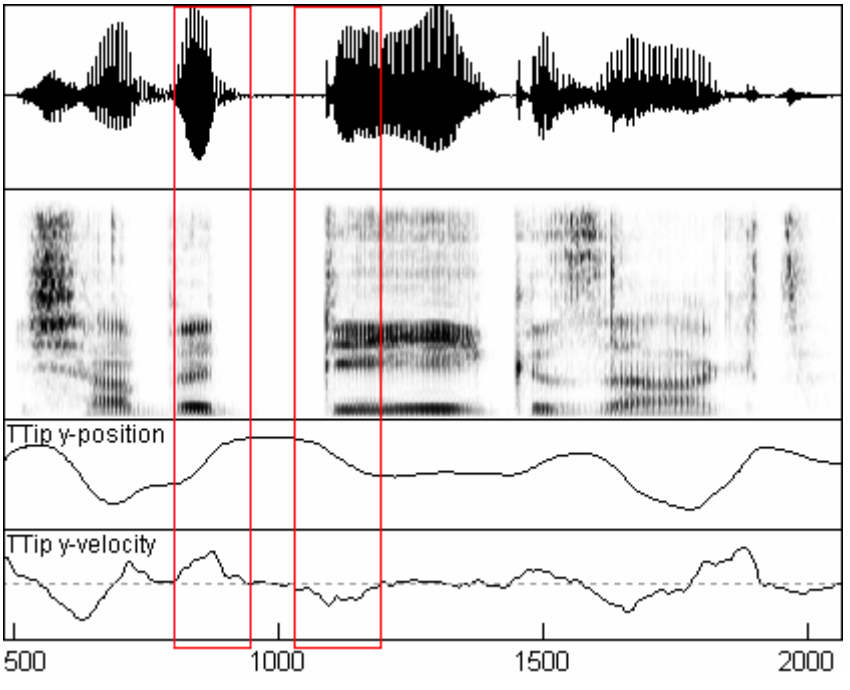

Fig. 1. The German sentence "Ich habe [de:] gesagt." in an oscillogram, spectrogram and two different EMA representations of the tongue tip movements. The upper EMA trajectory represents the y-position of the corresponding coil, (i.e. the tongue tip height) the lower trajectory shows the velocity in y-direction. The onset interval (left box) and the offset interval (right box) are marked. X-axis is time in milliseconds.

2 Method

The EMA session took place at the EMA lab of the Institute of Phonetics at the Ludwig-Maximilians-University Munich using the Carstens AG 500 articulograph with one subject. A total of ten receiver coils were used, four of them for the purpose

of correcting for head movement. Three coils were glued on the tongue to capture the oral constrictions performed by the tongue. To observe labial stops, both lips were also equipped with sensor coils. The sensor on the lower incisors was used to observe the jaw opening, while the sensor at the upper incisors as well as the sensors behind the ears and at the nasal bone are used as fixed points to remove the overlaid head movements from the articulator signals. See Table 1 for a list of coil locations and Fig 2 for a schematic view.

Table 1. List of EMA-receiver-coils used in this study, with their location and function

No.	coil location	function
1	tongue tip	tongue tip position
2	tongue mid	tongue mid position
3	tongue back	tongue back position
4	lower lip	lip protrusion
5	upper lip	lip protrusion/position
6	lower incisors	jaw opening
7	upper incisors	calibration
8	right ear	calibration
9	left ear	calibration
10	nasal bone	calibration

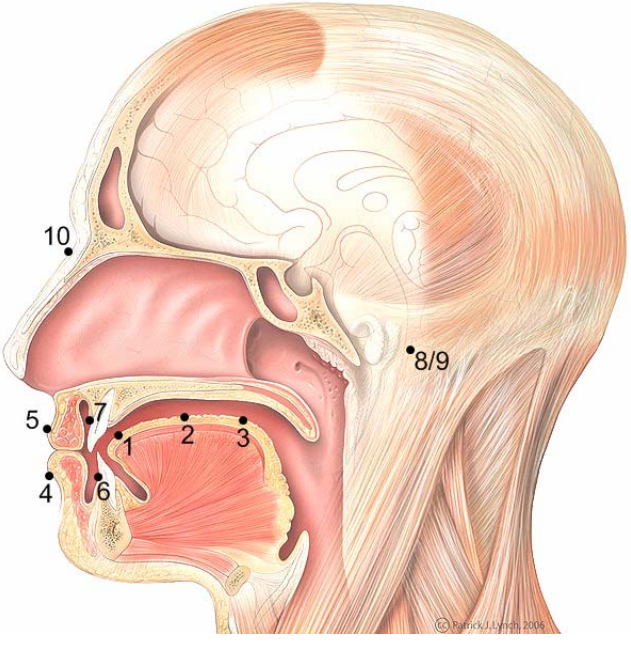

Fig. 2. Schematic view of the coil locations. See Table 1 for a description of coil locations and functions. (Drawing with permission of Patrick Lynch).

During the session we recorded a set of 90 plosive-vowel-syllables. Each syllable has been produced in the carrier sentence "Ich habe [xxx] gesagt." The plosives were [p],[b],[t],[d],[k],[g] the vowels were [i],[e],[a],[o],[u]. The 30 different syllables were recorded three times in total. The syllables were presented randomly to the speaker on a video-screen.

After normalization for head movement we identified the closure gesture in the articulatory trajectories and determined the duration of the onset and the offset using the software MVIEW [6].

3 Results

A. Onset Interval

The descriptive analysis of the consonantal onset showed that the duration ranges between 97ms to 166ms for apical plosives, 104ms to 197ms for dorsal plosives and 95ms to 157ms for labial plosives. Mean onset duration was 165ms for apical plosives, 132ms for dorsal plosives and 118ms for labial plosives (see Fig. 3). In the inferential analysis we tested if the factors 'place', 'voicing' and 'vowel' cause significant changes in onset and offset durations. 'Place' is divided into the specifications 'labial', 'apical' and 'dorsal', 'voice' is divided into "voiced and 'voiceless' and 'vowel' can be [i], [e], [a], [o] and [u].

The ANOVA showed a significant influence of the main factor 'place' and the interaction factor 'place+vowel' (see Table 2). A post-hoc Tukey-Kramer analysis indicated a highly significant difference in onset durations for labial and apical movements $p<0.0001$) and between apical and dorsal onsets ($p<0.0001$). There was a significant difference between dorsal and labial movements ($p<0.05$). Post hoc tests were performed for main effects only. The main factors 'vowel' and 'voice' were not significant in the ANOVA (Table 3).

Table 2. Results of the ANOVA for onset durations. (Significance levels: 0<***<0,001<**<0,01<*<0,05<.<0,1)

ONSET	Df	F	Pr(>F)	
place	2	39,0265	0,0000	***
voice	1	3,1214	0,0824	.
vowel	4	2,0219	0,1027	
place:voice	2	0,1629	0,8500	
place:vowel	8	2,8809	0,0088	**
voice:vowel	4	2,3249	0,0667	.
place:voice:vowel	8	2,1082	0,0488	*

B. Offset Interval

There is a strong variation in the offset duration. For the offset interval we observed durations between 110 and 285ms. The duration range for the plosive offset was 157ms to 285ms for apical plosives, 142ms to 265ms for dorsal plosives and 112ms to 227ms for labial plosives. The mean offset duration was 215ms for apical plosives, 198ms for dorsal plosives and 184ms for labial plosives (see Fig.4).

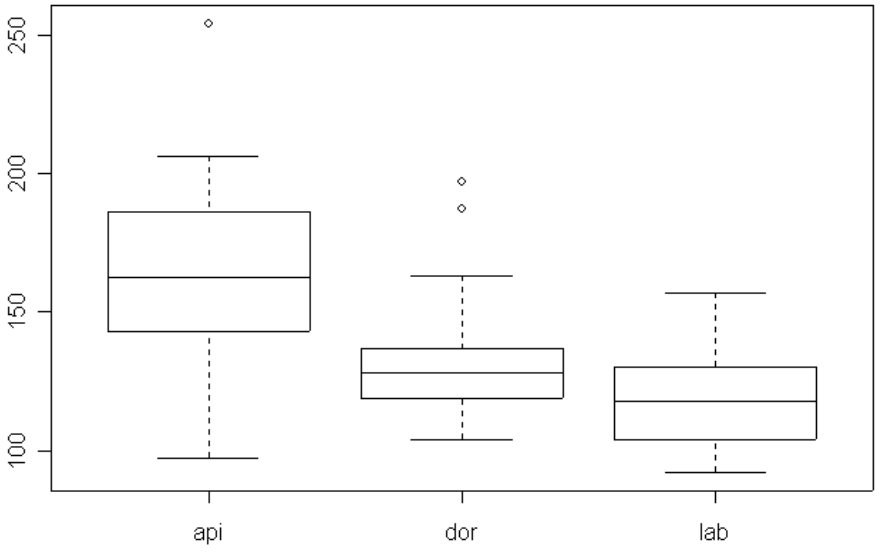

Fig. 3. Onset durations (in milliseconds) grouped by the factor 'place'. The dashed lines connect minimum and maximum values, the boxes represent the interquartile ranges and the horizontal lines inside the boxes represent mean values. Outliers are marked with rings.

Table 3. Results of the ANOVA for offset durations. (Significance levels: 0<***<0,001<**<0,01<*<0,05<.<0,1)

OFFSET	**Df**	**F**	**Pr(>F)**	
place	2	15,3654	0,0000	***
voice	1	0,5189	0,4741	
vowel	4	3,6006	0,0107	*
place:voice	2	7,9165	0,0009	***
place:vowel	8	2,9107	0,0082	**
voice:vowel	4	2,7922	0,0341	*
place:voice:vowel	8	1,5559	0,1576	

The ANOVA of offset durations showed a significant influence of the main factors 'place' and 'vowel' and the interaction factors 'place+vowel', 'place+vowel' and 'voice+vowel' (see Table 3). A post-hoc Tukey-Kramer analysis indicated a highly significant difference in onset durations for labial and apical movements $p<0.0001$). The post hoc analysis also showed statistically significant differences for [a] vs. [i] ($p<0.01$), [a] vs. [e] ($p<0.01$), [a] vs. [o] ($p<0.05$) and [a] vs. [u] ($p<0.05$). All these vowel combinations have a highly different degree of opening (see Fig. 5). The post hoc tests were performed for main effects only. The main factor 'voice' was not significant in the ANOVA.

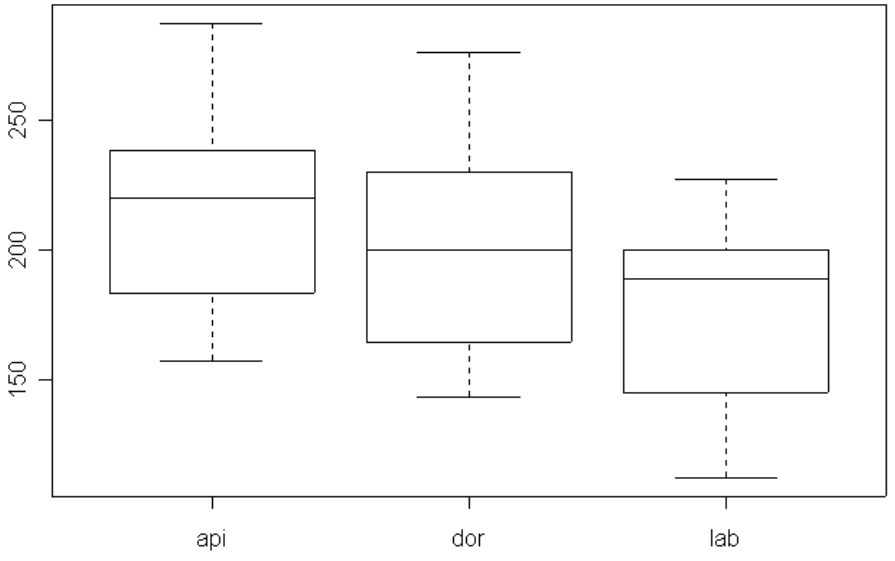

Fig. 4. Offset durations (in milliseconds) grouped by the factor 'place'. The dashed lines connect minimum and maximum values, the boxes represent interquartile ranges and the horizontal lines inside the boxes represent mean values.

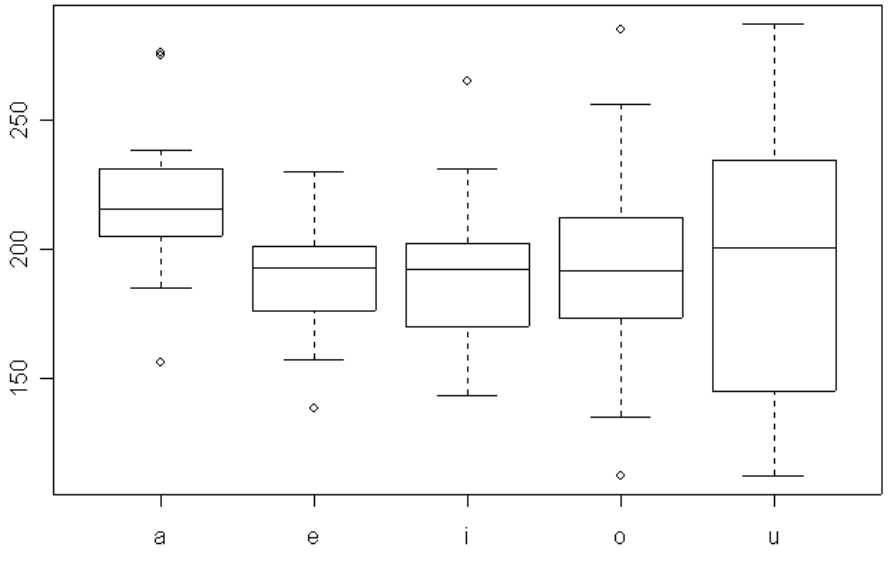

Fig. 5. Offset durations of consonantal stops in different vowel contexts. The dashed lines connect the minimum and the maximum values, the boxes represent the interquartile range and the horizontal lines inside the boxes represent the mean value. Outliers are marked with rings.

To compare the release velocity with the findings of Löfqvist, we also calculated the velocities for the offset interval. We observed a mean articulator speed of 11,8cm/s for labial movements, 12,5cm/s for tongue tip movements and 8,4cm/s for tongue body movements. The fastest labial release that we found was 18,6cm/s in the syllable [be:], the fastest tongue tip movement was 24,8cm/s in the release of the syllable [da:], the fastest dorsal opening was 18,5cm/s in the syllable [ga:]. The velocity values for labial movements are consistent with the findings of Löfqvist, where he analysed the lip kinematics [7].

4 Discussion and Conclusion

In contrast to other studies on articulator velocity and gesture durations [7],[8] our main interest was to make the values applicable in the field of articulatory speech synthesis.

We showed that there is a statistically significant difference in onset durations depending on the active articulator. The creation of synthetic action unit scores, which describe the articulatory movements in the vocal tract and serve as the input for an articulatory speech synthesizer, can now be enhanced by using different onset durations for different articulators. As an approximation we will use the mean durations we found in this study. The offset duration was also affected by the vowel [a] that increased the duration. This may be a result of the high degree of opening of the vowel [a].

The values of articulator velocity can be used to establish a plausibility check of the gestural scores to avoid speed values which are not natural. Since the speaking rate of the analysed speech was in a modal rate, we cannot determine if the articulator speed is affected by the global speaking rate. It is planned to perform an evaluation study after implementation of the results found in this study. We expect a higher degree of naturalness in the plosive burst and the aspiration noise in CV and CVC syllables.

Acknowledgments. We would like to thank Elizabeth Heller, Susanne Waltl and Manfred Pastaetter for their help during the EMA-Session and Mark Tiede for providing the software MVIEW. This work was supported in part by the German Research Council DFG grant KR 1439/13-1 and grant Kr 1439/15-1.

References

1. Birkholz, P., Kröger, B.J.: Vocal Tract Model Adaptation Using Magnetic Resonance Imaging. In: Proceedings of the 7th International Seminar on Speech Production, Belo Horizonte, Brazil, pp. 493–500 (2006)
2. Kröger, B.J., Birkholz, P.: A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)
3. Bauer, D., Kannampuzha, J., Kröger, B.J.: Articulatory speech re-synthesis: Profiting from natural acoustic speech data. In: Esposito, A., Vích, R. (eds.) Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions. LNCS, vol. 5641, pp. 344–355. Springer, Heidelberg (2009)

4. Kröger, B.J., Schröder, G., Opgen-Rhein, C.: A Gesture-Based Dynamic Model Describing Articulatory Movement Data. J. Acoust. Soc. Am. 98(4), 1878–1889 (1995)
5. Hoole, P., Nguyen, N.: Electromagnetic Articulography. In: Hardcastle, W.J., Hewlett, N. (eds.) Coarticulation – Theory, Data and Techniques, Cambridge Studies in Speech Science and Communication, pp. 260–269. Cambridge University Press, Cambridge (2000)
6. Tiede, M.: MVIEW: software for visualization and analysis of concurrently recorded movement data (2005)
7. Löfqvist, A.: Lip Kinematics in Long and Short Stop and Fricative Consonants. J. Acoust. Soc. A. 117(2), 858–878 (2005)
8. Adams, S.G., Weismer, G., Kent, R.D.: Speaking Rate and Speech Movement Velocity Profiles. Journal of Speech and Hearing Research 36, 41–54 (1993)