Intrinsic velocity differences of lip and jaw movements: preliminary results

Peter Birkholz[1], *Phil Hoole*[2]

[1]Clinic of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Aachen, Germany
[2]Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University Munich, Germany
pbirkholz@ukaachen.de, hoole@phonetik.uni-muenchen.de

Abstract

The observed kinematics of speech movements are the result of both the control by the brain and the biomechanical properties of the peripheral speech apparatus. For many kinematic phenomena, it is not clear whether they are actively controlled or intrinsic to the biomechanical system. This pilot study investigated the movement of sensors on the lips and the jaw in cyclical vowel transitions at specific speaking rates to identify possible intrinsic differences in the velocities of the articulators. Thereby, the lower lip was found to be significantly faster in approaching its targets than the upper lip, the mouth corners, and the jaw. Furthermore, for the mouth corners, backward movements were significantly faster than forward movements.

Index Terms: Speech kinematics, intrinsic velocity differences, articulatory control

1. Introduction

The observed movements of speech articulators result from the active control by the speaker as well as from the intrinsic biomechanical properties of the vocal tract [1]. The effect of active control on kinematic variables has been investigated for a wide range of factors, e.g., phonetic context, stress, and movement direction (e.g. [2, 3]). However, intrinsic relations between kinematic variables are less understood. Recently, the investigation of tongue back kinematics revealed substantial intrinsic differences in the velocity of forward, backward, upward, and downward movements [4]. The present study investigated potential intrinsic velocity differences in lip and jaw movements.

To find intrinsic relations experimentally, non-intrinsic factors that are known to affect kinematic variables, e.g., varying speaking rates and stress, must be well controlled. Our approach was to analyze the velocity of EMA-sensors on the jaw and lips during transitions between vocalic targets in V1-V2-V1 cycles as in [aʔoʔaʔoʔa ...] for different pairs of vowels. The utterances were spoken at controlled speaking rates and with a flat intonation to exclude these factors from affecting the movements. Furthermore, by using only vowel-vowel transitions instead of CV or VC transitions, kinematic influences due to the contact between the articulators and the vocal tract walls were minimized.

For the quantitative analysis of the lip and jaw kinematics, we used the parameters of a target-approximation model for articulatory movements. The measured movement trajectories were reproduced by optimizing the parameters of the model, and the parameters were analyzed to find potential intrinsic velocity differences. The target-approximation model is briefly introduced in Sec. 2, and Sec. 3 describes the data acquisition, the model-based reproduction of the measured movements, and the data analysis. The results and discussion follow in Sec. 4 and Sec. 5.

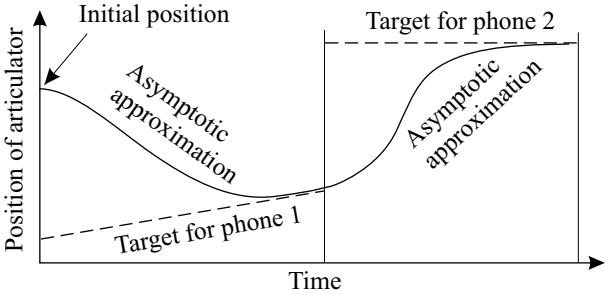

Figure 1: *A sketch of the target approximation model for supraglottal articulatory movements. The principle is the same as for the target approximation model for pitch targets [5].*

2. Target-approximation model for articulatory movements

The model to reproduce the observed trajectory of an articulator is based on the target approximation model in [6], which is in turn similar to the model for pitch movements in [5]. According to this model, the displacement of an articulator along a given spatial axis is the output of a linear dynamical system with a low-pass characteristic in response to a sequence of asymptotic targets. In this study, for a given utterance and articulator (sensor),

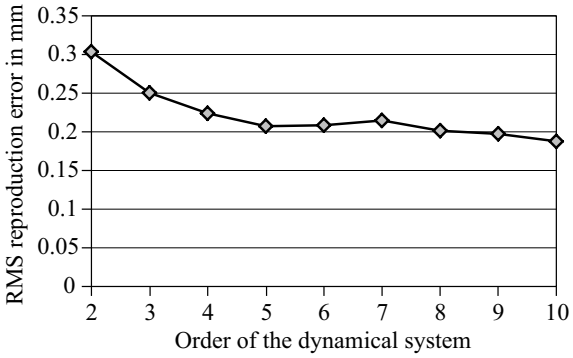

Figure 2: *RMS error for the reproduction of the measured trajectories of the tongue back sensor for the utterances of the corpus used in this study for different model orders of the dynamical system.*

one individual target is assumed for each vowel. Figure 1 illustrates the input (targets) and output (displacement curve) of the model for two successive targets. In contrast to [6], where targets were assumed to be static, they may change as a linear function of time here, as shown for the first target in Fig. 1. Hence, a spatial target position is allowed to change slowly over time at a constant velocity within the corresponding time interval. This is equivalent to the dynamic pitch targets assumed in [5] and proved here to allow much more precise reproductions of measured articulatory displacement curves. The dynamical system used for the approximation of a target is a cascade of multiple equal first order systems. Such a system was previously shown to account very well for the typical displacement-time and velocity-time functions of natural speech movements [7, 6]. It has the transfer function $H(s) = \frac{1}{(1+s\tau)^N}$, where s is the complex frequency, N is the order of the system (the number of cascaded first-order systems), and τ is the time constant of the system. The time constant is related to the time needed to approach a given target and inversely related to the velocity of the movement. τ is allowed to vary from one target to the next, i.e., each target can be approached with an individual velocity. At the border between targets, continuity of position, velocity, acceleration, and higher-order derivatives is preserved.

The order N of the system should be as low as possible for a short "reaction time" of the system when the target changes, but as high as necessary to reproduce the smooth shapes of natural movements. For $N = 2$, the system corresponds to a critically damped spring-mass system that has frequently been used to model articulatory movements (e.g. [8]). In this case, the step response of the system has a substantially shorter acceleration phase than deceleration phase, which contradicts observations of natural movements. For higher orders, the phases become progressively more symmetric and natural. Figure 2 shows, how the error between measured displacement curves and fitted model curves reduces, when the model order is increased. The error substantially drops up to an order of five or six and remains relatively stable for higher orders. Therefore, a 6th-order model was used in this study.

3. Method

3.1. Subject, corpus, instrumentation

One male native German speaker (32 years) produced a total of 40 V1-V2-V1-V2-... utterances of alternating vowels for the following ten pairs of German vowels: [i-ɛ], [i-a], [ɛ-a], [i-o], [i-u], [ɛ-o], [u-ɛ], [u-a], [a-o], and [u-o]. The sequence for each pair of vowels was spoken twice at each of two speaking rates: a fast rate with 375 ms per vowel and a slow rate with 500 ms per vowel. The speaking rate was regulated by a metronome. The utterances were spoken with a flat (neutral) intonation and a (partly reduced) glottal stop between the vowels, i.e., [aʔoʔaʔoʔ...]. Each utterance consisted of at least four transition cycles between the two vowels.

Articulatory movements were measured by means of the electromagnetic articulograph AG500 (Carstens Medizinelektronik) at a sampling rate of 200 Hz together with the audio signal. The articulograph recorded the 3D trajectories of small sensors attached to selected fleshpoints of the articulators in the fixed reference frame of the head. The movement data was low-pass filtered with a cut-off frequency of 20 Hz. The present study considered the 2D-sagittal movements of five sensors attached to the upper lip, lower lip, left mouth corner, right mouth corner, and jaw.

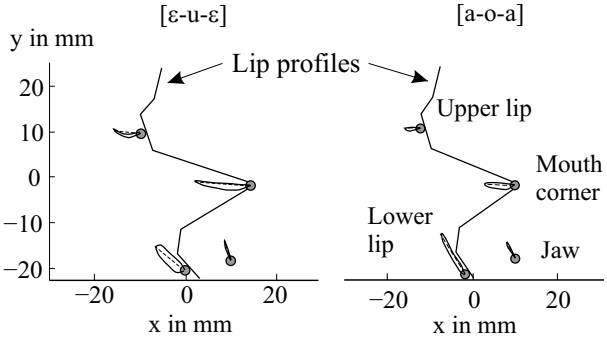

Figure 3: *Trajectories of the considered EMA sensors for the vowel cycles [ɛ-u-ɛ] (left) and [a-o-a] (right). The dashed lines are the first principal components of the trajectories, i.e., the main movement lines.*

3.2. Data preprocessing

To reduce nonsystematic variation in the measured displacement data, the 2D time signals of all sensors were first averaged over four consecutive V1-V2-V1 cycles to get the averaged 2D trajectory for one vowel cycle for

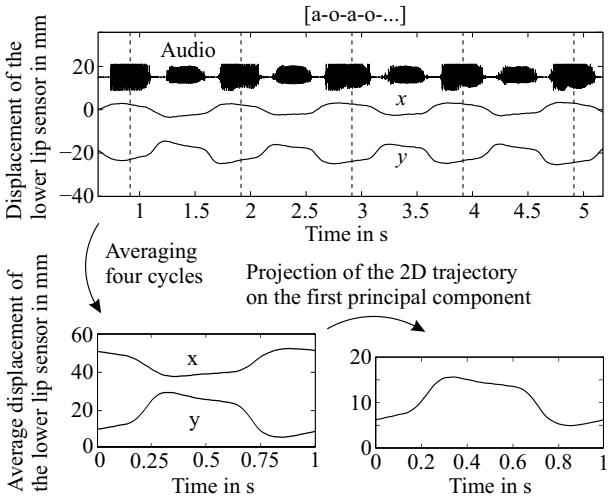

Figure 4: *Processing steps to get a 1D displacement signal for one vowel cycle from the 2D sensor trajectories using the example of the lower lip sensor in [a-o-a-o-...].*

each utterance and sensor. Figure 3 shows the averaged 2D trajectories for the considered sensors of two utterances. As these examples illustrate, the main direction of the cyclical trajectory of a given sensor was approximately the same across all utterances. The sensors for the upper lip and the mouth corners moved essentially along the horizontal axis, and the sensors for the jaw and the lower lip moved between a high-front and a low-back position. Therefore, the 2D trajectory of each utterance and sensor was reduced to a 1D displacement-time signal by projection of the 2D trajectory on the corresponding predominant movement line. The latter was determined as the first principal component of a given 2D trajectory. Trajectories with displacement ranges smaller than 2 mm were excluded from the analysis, because the estimation of model parameters (Sec. 3.3) was not reliable in these cases due to the measurement noise. Figure 4 illustrates the above processing steps for the lower lip sensor signals of the utterance [a-o-a-o-...] at the slow speaking rate.

To reduce the amount of data and variability, the trajectories of the left and right mouth corner sensors were averaged to a single mouth corner trajectory. Furthermore, we considered the trajectory of the lower lip *relative* to the jaw along with the measured *absolute* lower lip movements. Therefore, the jaw trajectory was subtracted from the lower lip trajectory. For a few utterances, we observed a slight drift of the sensor positions within an utterance. In these cases, the sensor position in the middle of the first V1 in a V1-V2-V1 cycle was not exactly the same as in the middle of the second V1. The drift was compensated by adding a linear function to the final displacement-time signal of the sensor that removed the difference between the first and last sample.

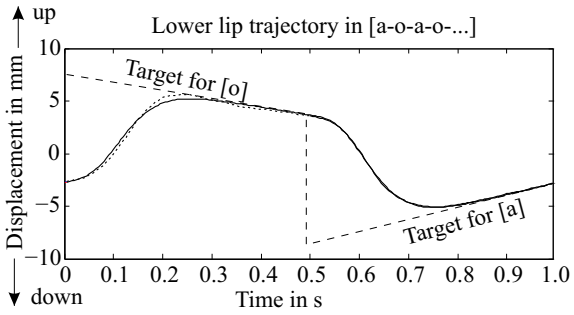

Figure 5: *Reproduction of the lower lip trajectory for the sequence [a-o-a-o-...] (slow speaking rate) with the target approximation model. The measured curve is drawn as dotted line and the fitted model curve as solid line. The optimized targets for the two vowels are shown as dashed lines.*

3.3. Estimation of model parameters

For each utterance and sensor, the 1D displacement curve was reproduced by the target-approximation model such that the RMS error between the measured curve and the model-fitted curve was minimized. For each model curve, the start time, position, slope, and time constant for each of the two vowel targets were jointly estimated by the Nelder-Mead simplex method [9] implemented in the Matlab toolbox version 7.4. To get close to the *global* minimum of the cost function, the optimization procedure was run 100 times for each curve with randomly varied initial values for the model parameters, and the best-fitting curve in terms of the RMS error was used for the analysis. Figure 5 shows an example for the measured and model-fitted curves for the utterance [a-o-a-o-...].

3.4. Analysis of model parameters

To detect potential velocity differences of the considered sensors, two-sample, two-tailed t-tests were performed with the time constants between all pairs of sensors for an overall significance level of $\alpha = 0.05$ (adjusted to $\alpha/4$ using Bonferonni correction). For each sensor, the time constants were pooled across the two speaking rates and movement directions.

Furthermore, to detect potential differences in the *direction-dependent* velocity of the sensors, paired two-tailed t-tests were performed between the samples of time constant estimates for the two movement directions of each sensor and speaking rate with $\alpha = 0.05$.

4. Results

Figure 6 shows the distributions of the estimated time constants for the movements ordered by sensor, speaking rate, and movement direction. For each sensor, the median value across both speaking rates and directions is written below the box plots. Hence, the upward-down-

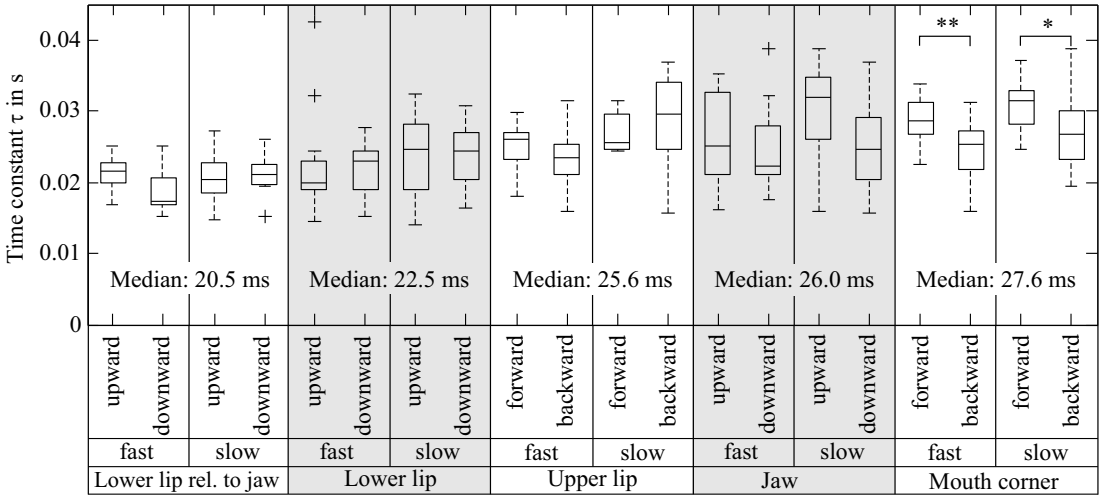

Figure 6: *Distributions of the time constants for the target-approximation movements for different sensors with respect to speaking rate and movement direction. The symbols * and ** mark, respectively, a significant difference ($p < 0.05$) and a highly significant difference ($p < 0.01$) between the means of the corresponding populations.*

ward movements of the lower lip relative to the jaw were generally the fastest (lowest time constant), while the forward-backward movements of the mouth corners were the slowest (highest time constant). According to the two-sample t-tests, the means of the time constant populations differed significantly between the lower lip relative to the jaw and all other sensors, and between the lower lip and all other sensors. Hence, the lower lip relative to the jaw was significantly faster than the absolute lower lip movement, and the absolute lower lip movement was faster than the movement of the upper lip, the jaw, and the mouth corners. The paired t-tests for *direction-dependent* differences of the time constants indicated significant differences for the mouth corner (cf. Fig. 6). Accordingly, backward movement of the mouth corners was generally faster than forward movement.

5. Discussion and conclusions

In this study, a new experimental design was used to discover potential *intrinsic* differences of the target-approximation velocities of sensors attached to the jaw and lips during speech movements in cyclical vowel transitions. For the considered subject and sensors, we found the target-approximation velocity of the lower lip to be significantly higher than the velocities of the upper lip, the jaw, and the mouth corners. Furthermore, backward movements of the mouth corners were significantly faster than forward movements. According to the experimental design, we hypothesize that the observed differences are truly intrinsic and as such are caused by the biomechanical properties of the involved muscles and tissues and not by motor control. We propose that such effects should be considered in the kinematic analysis of speech movements. Future studies with more subjects will show whether the observations are consistent across speakers or differ from speaker to speaker.

6. Acknowledgements

We thank Susanne Waltl and Lasse Bombien for their help with the acquisition of the EMA data, and Yi Xu and Santitham Prom-On for helpful discussions.

7. References

[1] S. Fuchs and P. Perrier, "On the complex nature of speech kinematics," *ZAS Papers in Linguistics*, vol. 42, pp. 137–165, 2005.

[2] C. L. Smith, C. P. Browman, and R. S. McGowan, "Extracting dynamic parameters from speech movement data," *Journal of the Acoustical Society of America*, vol. 93, no. 3, pp. 1580–1588, 1993.

[3] A. Parush, D. J. Ostry, and K. G. Munhall, "A kinematic study of lingual coarticulation in VCV sequences," *Journal of the Acoustical Society of America*, vol. 74, no. 4, pp. 1115–1125, 1983.

[4] P. Birkholz, P. Hoole, B. J. Kröger, and C. Neuschaefer-Rube, "Tongue body loops in vowel sequences," in *9th International Seminar on Speech Production (ISSP 2011)*, Montreal, Canada, 2011, pp. 203–210.

[5] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, 2009.

[6] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2011.

[7] K. Ogata and Y. Sonoda, "Reproduction of articulatory behavior based on the parameterization of articulatory movements," *Acoustical Science and Technology*, vol. 24, no. 6, pp. 403–405, 2003.

[8] J. A. S. Kelso, E. L. Saltzman, and B. Tuller, "The dynamical perspective on speech production: Data and theory," *Journal of Phonetics*, vol. 14, pp. 29–59, 1986.

[9] J. A. Nelder and R. A. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.