Model-Based Reproduction of Articulatory Trajectories for Consonant-Vowel Sequences

Peter Birkholz*, Bernd J. Kröger and Christiane Neuschaefer-Rube

Abstract

We present a novel quantitative model for the generation of articulatory trajectories based on the concept of sequential target approximation. The model was applied for the detailed reproduction of movements in repeated consonant-vowel syllables measured by electromagnetic articulography (EMA). The trajectories for the constrictor (lower lip, tongue tip, or tongue dorsum) and the jaw were reproduced. Thereby, we tested the following hypotheses about invariant properties of articulatory commands: (a) The target of the primary articulator for a consonant is invariant with respect to phonetic context, stress and speaking rate. (b) Vowel targets are invariant with respect to speaking rate and stress. (c) The onsets of articulatory commands for the jaw and the constrictor are synchronized. Our results in terms of high-quality matches between observed and model-generated trajectories support these hypotheses. The findings of this study can be applied to the development of control models for articulatory speech synthesis.

EDICS Category: SPE-SPRD

I. Introduction

A. Motivation and background

A model for the generation of articulatory movements is an important part of systems for articulatory speech synthesis. Such a model should be able to reproduce all real observable movements under different

The authors are with the Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital Aachen and RWTH Aachen University, Pauwelsstrasse 30, 52074 Aachen, Germany. E-mail (tel.): pbirkholz@ukaachen.de (++49 241 80 88956), bkroeger@ukaachen.de (++49 241 80 85222), cneuschaefer@ukaachen.de (++49 241 80 88942). Fax: ++49 241 80 82513

conditions, for example varying speaking rates and stress levels. It should have as few degrees of freedom as possible, but as many as necessary to explain observed movements. Ideally, the degrees of freedom should be related to the phonetic structure of an utterance. Currently, no such model exists. In the present study, we propose a dynamical model for articulation and we assess, how many degrees of freedom are needed for detailed reproductions of repeated consonant-vowel syllables. The corpus contained utterances at normal and slow speaking rate, stressed and unstressed syllables, and bilabial, alveolar, and dorsal consonants.

B. Related work

Many studies on the synthesis of articulatory trajectories are based on the idea that the movement of an articulator is realized by either the interpolation or the asymptotic approximation of a sequence of spatial target positions. Typically, one target per phoneme is assumed.

A triphone model for the generation of movements using interpolation between successive target positions was proposed by Okadome et al. [1]. In this model, the immediate neighbors of a phone in a sequence influence the target positions and velocities of the articulators of the current phone. Third-order polynomials are used to interpolate the articulatory trajectories between two phones. According to the authors, the model is restricted to the prediction of speech movements at a normal speaking rate. It cannot account for articulatory reduction that is observed when the speaking rate increases. Blackburn and Young [2] proposed a different model based on target interpolation. In this model, target positions are not fixed for a given context, but they are represented by probability density functions. In this way, it was tried to predict observed positional variability not only due to phonetic context, but also due to speaking rate. Standard linear interpolation between the targets was used to generate the trajectories.

In contrast to models that interpolate between targets, target approximation models define targets as asymptotic position values (e.g. [3], [4]). This category includes for example gesture-based speech production models [5]–[7]. Prom-on et al. [8] showed that not only supraglottal articulation, but also fundamental frequency curves can be effectively modeled by target approximation. Also the Equilibrium Point Hypothesis of motor control [9] resembles the idea of target approximation. The hypothesis suggests that movements arise from shifts in the equilibrium positions of the limbs or the speech articulators. In this framework, the equilibrium positions are the targets. Perrier et al. [10], [11] applied the hypothesis to speech motor control and proposed that piece-wise linear control signals for the equilibrium points underlie articulatory trajectories. Target approximation models are not only considered for speech production, but also for speech recognition [12]–[15]. They are used to impose constraints on the dynamics

of acoustic or articulatory features in an attempt to improve the performance of speech recognizers.

Target approximation models differ mainly with regard to the question, to what degree the targets are influenced by phonetic context and prosody. In his early study on vowel reduction, Lindblom [16] found asymptotic (acoustic) vowel targets that are invariant with respect to consonantal context and duration. The observed acoustic variability of vowels was explained as an undershoot of these targets when the vowel duration was too short to reach the asymptotic values. In contrast, Perkell et al. [17] found evidence that target specifications are modified by prosodic influences and reduction. In the same line, Wei et al. [18] and Dang et al. [19] proposed a carrier model of coarticulation, where targets are influenced by the phonetic context.

In contrast to target approximation models, Bouabana and Maeda [20] raised the idea to reproduce articulatory position trajectories using time-invariant linear second-order systems excited by impulse trains. However, the authors admitted that it was difficult to determine the number of impulses needed to adequately synthesize the observed movements, also because the impulses had no relation to the phonetic structure. In consequence, they proposed that excitations of the systems by rectangular time functions could be more appropriate than series of impulses. Rectangular time functions would be equivalent to sequences of spatial targets and therefore resemble the idea of target approximation. Ogata and Sonoda [21], [22] used impulse trains exciting linear systems to reproduce the velocity trajectories of articulators as opposed to position trajectories. This, as well, implicitly corresponds to the target approximation concept.

With regard to the location of targets, there is evidence that primary articulators of stop consonants have targets that lie beyond the positions that they can actually reach [23]–[26]. These targets are referred to as *virtual* targets. For example, the tongue tip target for [t] and the tongue dorsum target for [k] can be assumed somewhere above the palatal wall in the nose cavity. When a constrictor, starting from a vowel position, tries to reach its virtual target for a stop consonant, its velocity will be high at the time when it hits the vocal tract walls, and is then suddenly stopped by the collision. This is the typical pattern of natural constrictor movements observed in stop consonants [27]. In contrast, target-based transitions between successive vowels are rather smooth, because the articulators are not suddenly decelerated by collisions with the vocal tract walls. Using virtual targets, this difference between smooth and abrupt transitions can be elegantly modeled.

The dynamics of articulators in target approximation models are traditionally modeled by linear second-order systems in analogy to damped spring-mass systems [6], [7], [11], [28]. Usually, the systems are assumed to be critically damped to avoid an overshoot of the target positions. However, Kröger et

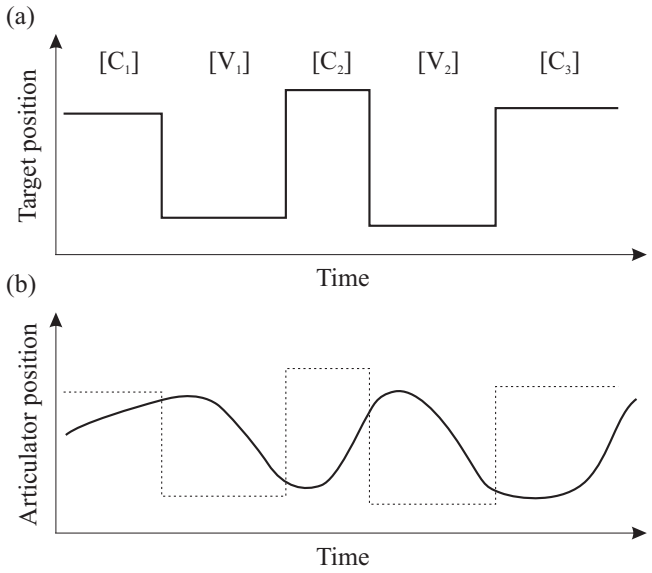


Fig. 1. (a) Sequence of target positions for an articulator with one target per phone. $[C_1]$, $[C_2]$, and $[C_3]$ are targets for consonants that generate closing gestures, and $[V_1]$ and $[V_2]$ are targets for vowels that generate opening gestures. (b) The articulatory trajectory produced by the target signal in (a).

al. [6] noted that natural articulatory trajectories cannot be fitted with high accuracy, when time-invariant critically damped second-order systems are excited with step functions. To achieve better fits, Kröger et al. introduced a force function that smoothly varies the system parameters during the interval of a target (gesture), making it a time-variant system. An alternative is to use linear transitions between the target positions instead of stepwise changes, as by Perrier et al. [11]. The drawback of both methods is that they introduce additional degrees of freedom in the movement specification. In contrast, Ogata and Sonoda [21], [22] proposed to model articulatory dynamics with higher-order dynamical systems, inspired by Milsum [29]. They allow to reproduce certain articulatory trajectories with high accuracy and few degrees of freedom in the movement specification.

In this study, we combine the idea of target approximation with an effective dynamical system for articulators to generate detailed reproductions of observed movements.

II. Model description

Before we go into mathematical details, we start with a short overview of the model. Basically, the trajectory of an articulator along a spatial coordinate axis is assumed to be the output of a time-variant dynamical system. The input to the system is a sequence of asymptotic target positions, where each phone or gesture defines one target for a given time slice. Thus, the input signal is divided up into

sections for discrete phonetic units. The boundaries between the sections are characterized by stepwise changes of the target position. Figure 1 (a) illustrates a possible input function for successive opening and closing gestures of an articulator. The dynamical system for an articulator has the effect of a low-pass filter. Therefore, the stepwise changes between the targets for different phones are translated into smooth changes of the actual articulator positions. Figure 1 (b) shows the articulator position signal for the input signal depicted in Fig. 1 (a). The dynamical system has one parameter, the time constant τ, that controls how fast the system output approximates the input. In the proposed model, the time constant is assumed to remain constant within the time slice for a phone, but is allowed to vary from one time slice to the next. Hence, the movement of an articulator towards a target is specified by the target position and the time constant. The target and the associated time constant that control the realization of a phone or gesture will be referred to as *articulatory command*. Each command has an onset time where it starts to take control of an articulator. It keeps the control until the next command begins.

The described ideas resemble the target approximation model by Xu [3], who applied it for the reproduction of F_0 contours and motivated its application for the control of supraglottal articulation. In the following, we present a quantitative formulation of the model. The dynamical system we propose in this framework is based on the studies by Ogata and Sonoda [21], [22]. The basic idea is to describe the dynamics of an articulator by a cascade of several identical first-order linear systems. The transfer function of such a system is

$$H(s) = \frac{Y(s)}{X(s)} = \frac{1}{(1 + s\tau)^N}, \tag{1}$$

where s is the complex frequency, τ is the time constant, and N is the order of the system. In agreement with Ogata and Sonoda [21], we set $N = 10$. A tenth-order system reproduces very well the bell-shaped velocity profiles observed in natural directed movements for step functions as input. For lower orders, the velocity profiles would become progressively asymmetrical and would not allow detailed reproductions of observed movements. For higher-order models, the delay between the input and output signals would become very high. A tenth-order model has a reaction time of roughly 50–100 ms, which corresponds to measured delays between the onsets of muscle activity and articulatory motion [30], [31].

The only free parameter of the system defined by Eq. (1) is the time constant τ. For the proposed model, we assume that τ remains constant during a command, but may vary between one command and another. Hence, the dynamical system is time-invariant during one command, but not across the boundaries between commands.

We now derive the time-function $y(t)$ of an articulator position within the time slice of a command.

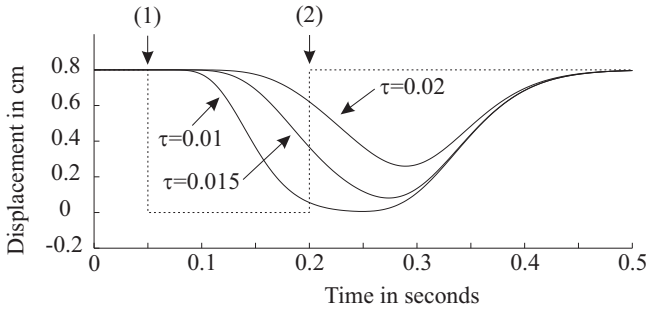


Fig. 2. Response of the proposed dynamical system to a sequence of two articulatory commands that control the vertical position of the jaw, for example. The onsets of the commands are marked with (1) and (2). The targets associated with the commands are drawn as horizontal dotted lines. The two commands can be interpreted as an opening gesture followed by a closing gesture. The solid lines show the resulting articulatory trajectories for three different time constants of the *first* command.

$y(t)$ is the response of the system (1) to an input signal $x(t)$. To obtain the input-output relations in the time domain, we rewrite Eq. (1) as

$$Y(s)[(1+s\tau)^N] = X(s) \tag{2}$$

and apply the correspondences

$$\begin{aligned} X(s) &\quad \bullet\!\!-\!\!\circ \quad x(t) \\ s^i Y(s) &\quad \bullet\!\!-\!\!\circ \quad \mathrm{d}^i/\mathrm{d}t^i \; y(t). \end{aligned}$$

This results in the following differential equation for $y(t)$:

$$\binom{N}{0}\tau^N y^{(N)} + \binom{N}{1}\tau^{N-1} y^{(N-1)} + \ldots + \binom{N}{N} y^{(0)} = x. \tag{3}$$

Here, $\binom{n}{k}$ denotes the binomial coefficient and $y^{(i)}$ the ith derivative of y with respect to time. Solving the equation for $y(t)$ yields

$$y(t) = (c_0 + c_1 t + \ldots + c_{N-1} t^{N-1}) e^{-t/\tau} + b, \tag{4}$$

where $b = x(t)$ is the constant input signal (target position) of the command. The coefficients c_i depend on the initial conditions at the onset of the command. They result from the continuity constraints that we require for $y(t)$ and its $N-1$ derivatives at the boundary between two commands. Hence, $y(t)$ and its derivatives at the offset of a command determine the coefficients c_i for the next command.

The required equations are summarized in the following. Let $M = N - 1$ and $a = -1/\tau$. Then, the n-th derivative of $y(t)$ can be written as

$$y^{(n)}(t) = e^{at} \sum_{k=0}^{n} \left[a^{n-k} \binom{n}{k} \left(\sum_{i=k}^{M} \frac{i!}{(i-k)!} c_i t^{i-k} \right) \right] \tag{5}$$

for $n = 1 \dots M$. Re-sorting the terms in Eq. (5) yields

$$y^{(n)}(t) = e^{at} \sum_{i=0}^{M} t^i q_i^{(n)} \tag{6}$$

with

$$q_i^{(n)} = \sum_{k=0}^{\min\{M-i,n\}} a^{n-k} \binom{n}{k} c_{i+k} \frac{(k+i)!}{i!}. \tag{7}$$

Equations (4) and (6) can be used to determine $y(t)$ and the derivatives $y^{(1)}(t) \dots y^{(M)}(t)$ at the offset of a command. These values are taken as the start values for $y(t)$ and its derivatives in the next command interval. In this way, the system state at the offset of one command is transferred to the next. When we assume that the local time starts with $t = 0$ in each command interval, the coefficients $c_0 \dots c_M$ for Eq. (4) can be calculated as

$$\begin{aligned} c_0 &= y(0) - b \\ c_n &= \left(y^{(n)}(0) - \sum_{i=0}^{n-1} c_i a^{n-i} \binom{n}{i} i! \right) / n! \end{aligned}$$

for $n = 1 \dots M$.

For a quantitative example of the proposed model, consider the control of the vertical jaw position depicted in Fig. 2. The input to the model are two articulatory commands for the realization of an opening gesture and a closing gesture. The targets associated with the commands are drawn as horizontal dotted lines, and the onsets of the commands are marked with (1) and (2). The solid lines are the articulatory trajectories, i.e. $y(t)$, for three different values for the time constant of the first command. The initial position of the articulator and the time constant of the second command is equal for all three cases. The trajectories highlight the following properties of the model: the delay between the onset time of a command and the resulting movement; the inverse relation between the time constant and the articulator velocity; the undershoot of a target depending on the command duration and the time constant.

Up to here, the model is well suited to reproduce the smooth articulatory transitions between vowel targets. However, as discussed before, the constrictor trajectory in the vicinity of a stop consonant is usually not smooth. For example, when the tongue tip approaches the (virtual) target for an alveolar stop consonant, it will be suddenly decelerated when it hits the vocal tract walls at some position y_0 on its way. To model this collision behavior, $y(t)$ is set to y_0 during the closure interval. This clipping

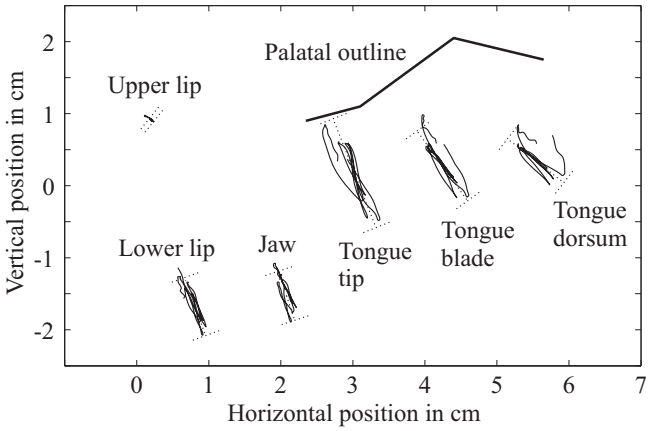


Fig. 3. Two-dimensional trajectories of the transducer coils for the sequence [ˌnanaˈnana] at normal speaking rate. The first principal component of each trajectory is drawn as dotted line. The length of each line is four times the standard deviation of the data along the corresponding principal component. The thick black line shows a part of the palatal outline.

introduces points of discontinuity in the trajectory of the constrictor at the onset and offset of the closure, rendering the model non-linear.

In the present study, the onset and offset times of consonantal closures were estimated from the speech waveforms, which were recorded together with the articulatory trajectories. The onset and offset of consonantal closures typically results in distinct changes in the speech waveform that can be easily identified. The measured positions of a primary articulator at these points in time was used to determine y_0. However, when the proposed model is applied to control articulators in the framework of a vocal tract model, the clipping of the trajectories must be handled by collision detection between the articulators and the vocal tract walls.

III. Experiments

The articulatory commands introduced above cannot be directly observed. However, using an analysis-by-synthesis approach, command sequences can be found that produce model-based articulator trajectories that closely match measured trajectories. This section presents two analysis-by-synthesis experiments for different presumptions about the command parameters.

Initial experiments showed that different command sequences can produce equally detailed fits of measured trajectories, when all command parameters can be varied independently. A major cause for this non-uniqueness of the solutions is the inter-relation between the target position and the time constant of a command. Within certain limits, the effect of a change of the gestural target for an articulator on the

trajectory can be compensated by a change of the time constant. For example, the slope or peak velocity of an opening gesture for some articulator can be kept constant when the target position is lowered and the time constant is increased correspondingly. This basically means that the proposed model has more degrees of freedom than needed to reproduce observed articulatory trajectories. In the two experiments described below, we tested different assumptions about the articulatory commands that effectively reduced the degrees of freedom.

The trajectories that we chose for the model-based reproductions were that of the jaw and the constrictor (tongue tip, tongue dorsum, lower lip) in repeated consonant-vowel syllables. In this way, we tested the model for the reproduction of trajectories with discontinuities, i.e. constrictor trajectories, and those without discontinuities, i.e. jaw trajectories.

A. Data

One female subject produced one [ˌCVCVˈCVCV]-sequence (C=consonant, V=vowel) for each combination of the vowels {/a/, /ɛ/, /e/, /o/, /ɔ/, /ø/, /œ/} with the consonants {/m/, /n/, /ŋ/} at both normal and slow speaking rate [32]. The slow sequences were produced especially clear (hyperarticulated). All sequences were spoken with secondary stress on the first syllable and primary stress on the third syllable.

Articulatory trajectories were recorded by means of electromagnetic articulography at a sampling rate of 200 Hz (EMA, AG100, Carstens Medizinelektronik 2002). A Kaiser-window low-pass FIR filter with a passband from 0 to 50 Hz, a transition bandwidth of 20 Hz, and a stopband attenuation of 50 dB was used to reduce the measurement noise in the trajectories. The current study considered the two-dimensional trajectories of transducer coils attached to the upper lip, lower lip, lower jaw (below the lower incisors), tongue tip and tongue dorsum. In synchrony with the kinematic measurements, the acoustic signal was recorded at a sampling rate of 16 kHz. The recordings of [ˌnønøˈnønø] and [ˌnœnœˈnœnœ] at the slow speaking rate were incomplete and therefore excluded from the experiments.

B. Preprocessing and Labeling

In each sequence, we marked the beginning and the end of the closure phase of each consonant by visual inspection of the acoustic signal. The waveform parts for the nasal consonants could be well identified and distinguished from the adjacent vowels. Furthermore, we defined the time interval to be used for the analysis and trajectory reproductions in each sequence. The beginning of the interval was set to the end of the first consonantal closure, and the end of the interval was set roughly to the middle

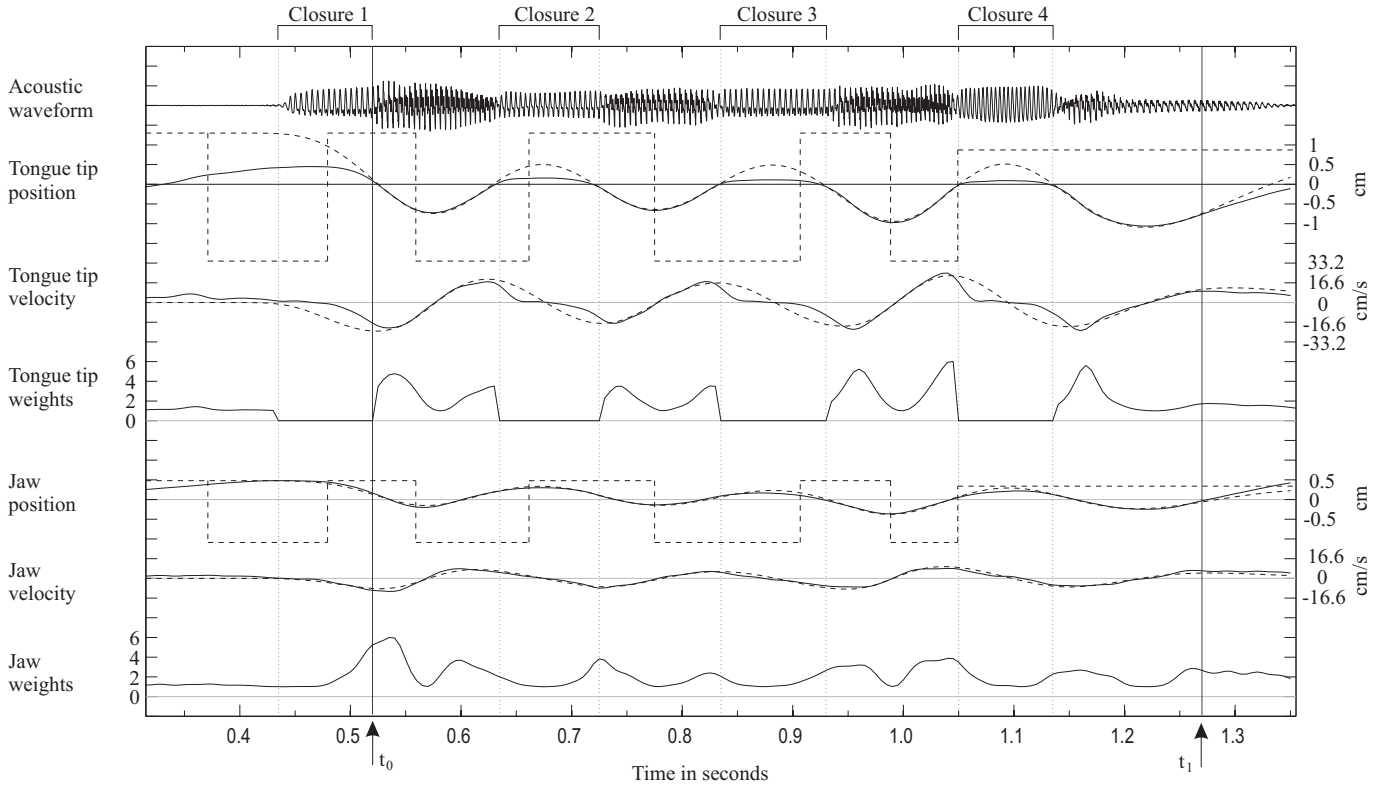


Fig. 4. Time-signals of the positions, velocities, and optimization weights (cf. Eq. 9) of the tongue tip and the jaw in the sequence [ˌnanaˈnana] at normal speaking rate. The dashed rectangular time functions show the targets for the opening and closing gestures for both articulators. The smooth solid lines show the recorded position and velocity signals of the tongue tip and jaw. The smooth dashed lines show their model-based reproductions. During the closure intervals, the tongue tip of the recorded signal remains approximately at the same position because it is braced against the teeth ridge. In contrast, the output of the dynamical system is shown without the consideration of clipping during palatal contact and therefore overshoots the closure position at 0 cm. Note the delay of over 100 ms between the onsets of commands for the closing gestures (high targets) and the onsets of the corresponding closure intervals.

of the final vowel. Figure 4 depicts the closure intervals and the analysis interval (from t_0 to t_1) for the sequence [ˌnanaˈnana].

The two-dimensional trajectories of the EMA coils for the sequence [ˌnanaˈnana] between t_0 and t_1 are shown in Fig. 3. It can be seen that the articulators move roughly along the same straight lines during the repeated opening and closing movements. Therefore, the movement of each articulator in each sequence was reduced to one dimension by the projection of the two-dimensional trajectories on the first principle component in the data between t_0 and t_1. The principal components are drawn as dotted lines in Fig. 3. For sequences with the consonant [m], the one-dimensional trajectories for the upper lip, the lower lip,

and the jaw were extracted. For sequences with [n] and [ŋ], the trajectories of the jaw and the tongue tip or the tongue dorsum were considered, respectively. The one-dimensional trajectories of the tongue tip and the jaw in the sequence [ˌnanaˈnana] are depicted in Fig. 4. For the tongue tip curve, the zero-line defines the positions where the closure intervals of the consonants begin and end.

C. Curve Fitting

The aim of our experiments was to estimate the articulatory commands underlying the observed trajectories for the constrictor and the jaw in each sequence, i.e. the commands that minimize the difference between the observed and model-based trajectories. To find the optimal command parameters, we used the Nelder-Mead simplex method [33] implemented in the Matlab toolbox version 7.4. This method finds the minimum of a scalar objective function of several variables, starting at an initial estimate. In the following, the variables and the objective function will be described. The initial estimates for the variables will be discussed later.

In the context of our model, one command per phone and articulator is assumed. Hence, 16 commands were required for the eight phones and the two articulators considered in a [ˌCVCVˈCVCV] sequence. One additional command per articulator was appended to the end of each sequence to control its movement towards the final rest position. Therefore, 18 commands per sequence had to be specified. Each command is defined by three parameters: the target position, the time constant, and its onset time. For 18 commands, this yields 54 parameters. In the experiments below, not all of the parameters will be optimized independently. Instead, groups of two or more parameters, that are assumed to be equal, may be represented by one variable in the optimization.

The objective function to be minimized during the optimization was designed to represent the dissimilarity between the observed trajectories and the model-based trajectories of a sequence. It was defined as

$$E = \sqrt{\left[\sum_i w_i (y_i - \widetilde{y}_i)^2\right] / \sum_i w_i} \tag{8}$$

with

$$w_i = 1 + a \cdot v_i^2 / v_{\max}^2, \tag{9}$$

where i is the sample index, y_i the original signal, $\widetilde{y}_i$ the model signal, w_i the weight signal, $v_i = (y_i - y_{i-1})/T_s$ the velocity signal, $T_s = 200$ Hz the sampling rate, and $v_{\max}$ the maximum velocity of the curve under consideration. All samples between t_0 and t_1 of the constrictor trajectory and the jaw trajectory were considered. If w_i would equal 1 for all i, Eq. (8) would become the simple root mean

square function. However, preliminary investigations revealed that this would underestimate the parts of the signal with high velocities, especially just before and after closures. To improve the match of these signal parts at a slight expense of the match during stationary parts, the weights w_i were introduced. The factor a in Eq. (9) defines the relative importance of position vs. velocity for the match and was set to $a = 5$ in our simulations. For constrictor trajectories, w_i was set to zero for all samples within the closure intervals to exclude them from the optimization.

Figure 4 illustrates the optimization results for the tongue tip and the jaw trajectories of the utterance [ˌnanaˈnana]. The optimized target sequences for both articulators are drawn as dashed rectangular functions. Observed trajectories are drawn as solid lines, and model-based trajectories as dashed lines. The model-based reproduction of the curves appears to be very good. Note that during the closure intervals, the model-based tongue tip positions overshoot the closure plateaus of the original tongue tip trajectories towards the virtual targets. As discussed before, the *actual* model-based articulator position is assumed to be clipped during these intervals.

D. First experiment

Initial optimization experiments indicated that mainly two factors determine the quality of curve fitting, namely, the number of independent variables to be optimized and the initial estimates for the variables. Ideally, the number of variables should equal the inherent degrees of freedom of the problem. Too many variables make the optimizer prone to get stuck in a local minimum of the objective function, and too few variables can make it impossible to achieve a good solution at all. When all command parameters of a sequence are optimized simultaneously, the solution strongly depends on the initial estimates. Therefore, to reduce the number of independent variables, we made the following assumptions:

- Jaw and constrictor commands for the same phone in a sequence start at the same time. This assumption of synchrony is corroborated by the idea that these articulators move as a coordinated structure [34].
- Each articulator has one common asymptotic target for each phoneme in the same context and for the same speaking rate. For the sequence [ˌnanaˈnana] at normal speaking rate, for example, we thus assume that there is one target position for both the tongue tip and the jaw for all [n], and another pair of target positions for all [a].

The number of variables optimized for each sequence in this experiment was thus 31 (9 command onset times + 4 target values + 18 time constants). Besides the above two assumptions, we considered requiring the same time constant for the constrictor and the jaw for the same gesture. However, with this

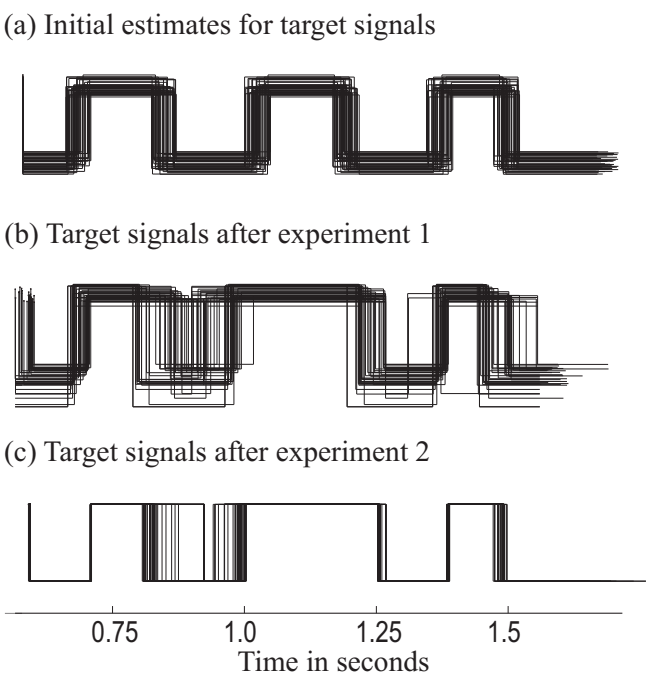


Fig. 5. Variation of the target signals for the tongue tip in [ˌnɔnɔˈnɔnɔ] after the optimization of 50 initial target signal estimates in experiment 1 and experiment 2.

strict constraint, a detailed fit could not be achieved for most sequences. Therefore, the time constants of all commands were optimized independently. Interestingly, the resulting time constants were nevertheless strongly correlated between the jaw and the constrictor targets, as shown in the next section.

As mentioned above, the result of an optimization with many variables depends on the initial estimates. To increase the chance to get close to the optimal solution, we made 50 optimization runs for each sequence with varying initial values. Before each run, the time constants of all commands were set to random values between 0.01 and 0.02 seconds. The position of consonant targets was set to a random number between MAX and MAX+1 cm, where MAX is the highest value of the corresponding observed trajectory. Accordingly, the position of vowel targets was set to a random number between MIN and MIN-1 cm, where MIN is the lowest value of the corresponding observed trajectory. The onsets of the commands were set randomly between $50 \ldots 100$ ms before the midpoints of the corresponding acoustic segments. Figure 5 (a) illustrates the variation of the 50 target signals for the tongue tip in [ˌnɔnɔˈnɔnɔ] that served as initial estimates for the optimization.

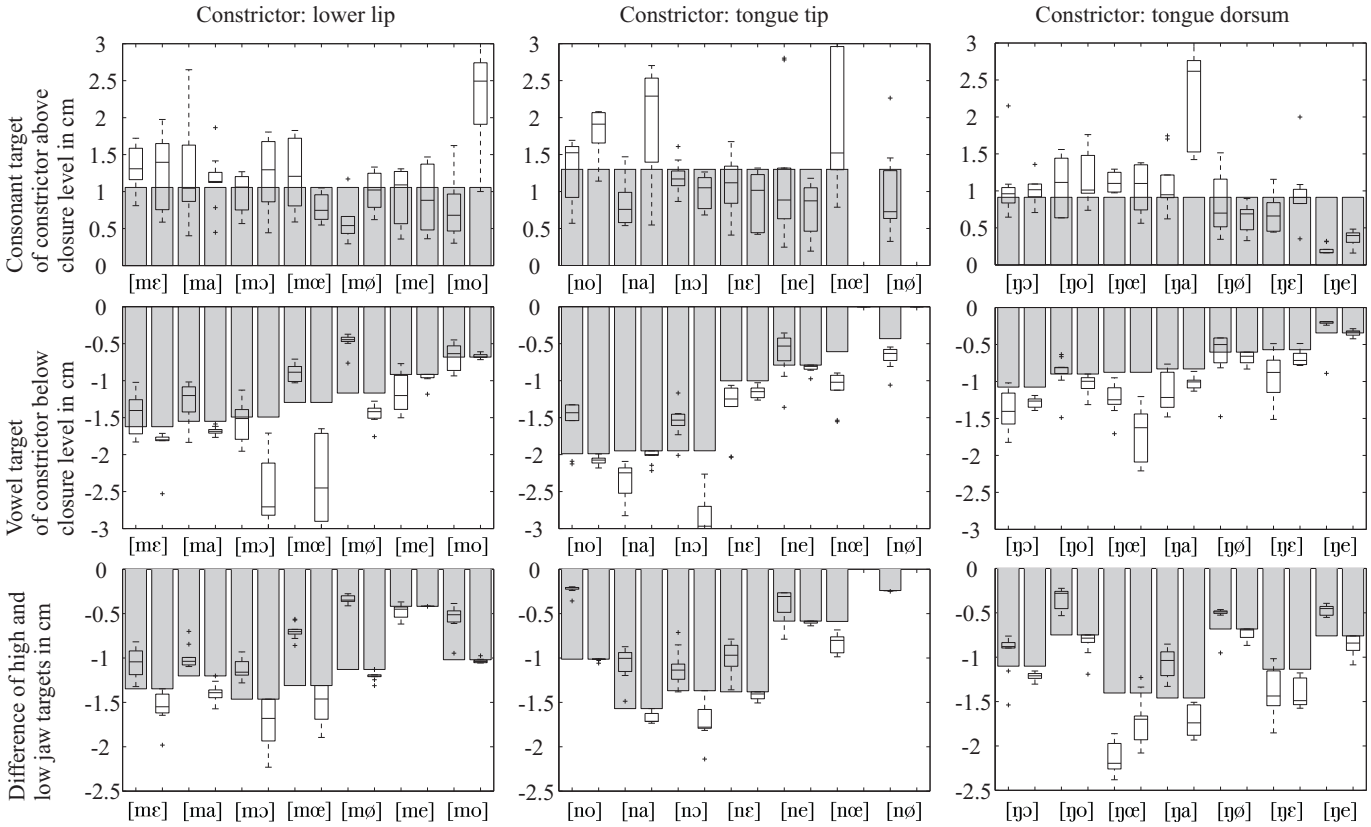


Fig. 6. Results for the targets of experiment 1. The box plots show the targets for the constrictor for closing gestures (top row), for opening gestures (middle row), and the difference between the targets for opening and closing gestures of the jaw (bottom row) for the 10 best optimization runs. The left, middle, and right columns display the results for sequences with the consonants [m], [n], and [ŋ], respectively. Two box plots are shown for each consonant-vowel combination next to each other: one for the normal (left) and one for the slow (right) speaking rate. The gray bars display the common median value of the constrictor targets (top row), the lowest position in the constrictor trajectory for each CV-combination (middle row), and the difference between the maxima and minima of the jaw trajectory for each CV-combination (bottom row).

E. Results of first experiment

For all utterances, the optimization results of the 50 runs with different initial estimates varied rather strongly. Figure 5 (b) illustrates the variations in terms of the 50 optimized target signals for the sequence [ˌnɔnɔˈnɔnɔ]. The variation appears somewhat greater than for the initial estimates in Fig. 5 (a). Especially in the region around the command for the second opening gesture, there is a great deal of variation of the command onset times. However, the error between the observed and model-based trajectories was generally equally low for most of the 50 runs. Hence, with the given degrees of freedom, different sets of variable values can produce equally good approximations of the original trajectories. Therefore, we

further reduced the number of independent variables in experiment 2 described below.

Figure 6 summarizes the results for the targets found in the first experiment. For each sequence, the distribution of three quantities is displayed for the ten best results of the 50 optimization runs: the (virtual) consonant target of the constrictor, the vowel target of the constrictor, and the difference between the low and high targets for the jaw. Two box plots are shown for each consonant-vowel combination – one for the normal and one for the slow speaking rate sequence.

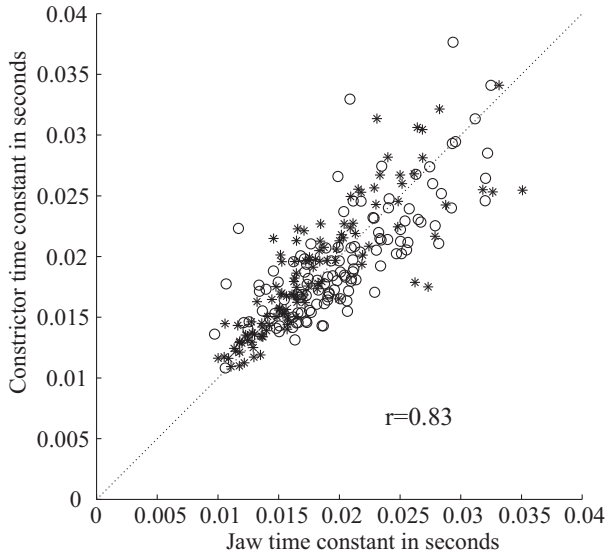


Fig. 7. Results for time constants after experiment 1. For all phones in all sequences, the time constant of the constrictor command is plotted against the time constant of the corresponding jaw command. Time constants of commands for opening gestures are displayed as circles and for closing gestures as stars.

An overview of the time constants for the commands found in the first experiment is given in Fig. 7. For each phone in each sequence, the constrictor time constant is plotted against the jaw time constant. A regression coefficient of 0.83 confirms that the time constants for both articulators are strongly coupled. When time constants for opening gestures (circles in Fig. 7) and closing gestures (stars in Fig. 7) are analyzed separately, an interesting trend can be observed. For commands for opening gestures, the jaw time constant is greater than the constrictor time constant for 64.2% of the vowels. Hence, the jaw opens somewhat slower than the constrictor in the majority of cases. Conversely, the jaw time constant is less than the constrictor time constant for 70.0% of the commands for closing gestures. This implies that the jaw moves up somewhat faster than the constrictor in 70.0% of the cases.

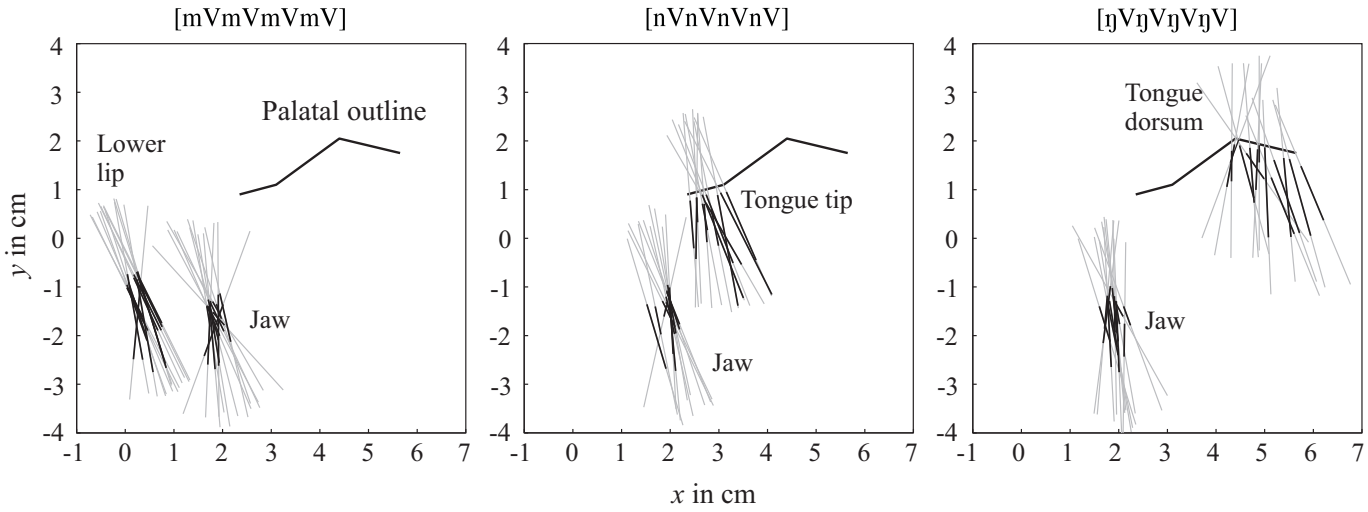


Fig. 8. The gray lines show the main movement directions (first principal components) of the trajectories of the constrictor and the jaw for the sequences with the consonant [m] (left), [n] (middle), and [ŋ] (right) in the midsagittal plane. One gray line is shown for each vowel in combination with the consonants for both the normal and slow speaking rates. The black parts of the lines range from the minimum to the maximum excursion of the corresponding articulator.

F. Second experiment

The results of the first experiment indicated that some of the variables considered for optimization were redundant, because equally detailed model-based approximations of the observed movement trajectories could be obtained with quite different sets of command parameter values. A second experiment was conducted to test the performance of the model with further reduced degrees of freedom.

One idea was to set the constrictor targets to the same predefined position for all sequences with the same consonant. If our targets were two-dimensional points in the midsagittal plane, this would mean to presume the same virtual 2D-target position for the constrictor in a consonant independent of vowel context, stress, and speaking rate [26]. If this were the case, the linear extensions of the main movement vectors in consonant-vowel sequences with the same consonant and different vowels would cross at this virtual position in the midsagittal plane. To test this hypothesis, the extended movement vectors (principal components) of the constrictor and jaw were plotted in Figure 8 for different context vowels. It is evident that the lines do *not* cross in exactly the same point for either articulator. However, especially for the tongue tip and the tongue dorsum, the lines clearly *converge* in a region above the palate. The constant *one-dimensional* target positions proposed here would fit quite well in these convergence regions. For estimates of the constant target positions, we drew upon the results of experiment 1. For each constrictor,

the median value of the consonant targets found in the ten best optimization runs of all corresponding sequences was taken as the common target. These target positions are 1.1 cm (above the closure position) for the lips, 1.3 cm for the tongue tip, and 0.9 cm for the tongue dorsum. In the top row of Fig. 6, these values are displayed as gray bars. With a few exceptions, these values lie between the minima and maxima of the ten best results found for these targets in experiment 1.

To further reduce the number of variables, we picked up the hypothesis by Lindblom [16], that there exist asymptotic vowel target positions, which are independent of consonantal context and duration. In line with this hypothesis, we now assumed the same fixed vowel targets for the constrictors independent of speaking rate. These vowel targets are most likely achieved at a slow speaking rate in stressed syllables. Therefore, for each consonant-vowel combination, we preset this target to the lowest value of the constrictor trajectory of the sequence recorded at the slow speaking rate. The positions of these targets (below the consonantal closure points) are indicated by the gray bars in the second row of Fig. 6.

In line with the above arguments, we also assumed a fixed asymptotic consonant target and a fixed asymptotic vowel target for the jaw movement of each consonant-vowel combination independent of speaking rate. These targets were estimated as the maxima and minima found in the corresponding jaw trajectories. The common differences between the high and low jaw targets for the normal and slow speaking rate are depicted by the gray bars in the bottom row of Fig. 6.

After defining all targets in all sequences, the only variables that remained to be optimized for each sequence were the time constants of all commands and common onset times of commands for the same phone. These are 27 variables (9 command onset times + 18 time constants). As in experiment 1, we made 50 optimization runs with randomized initial values to assess the variation of the solutions.

G. Results of second experiment

Despite the reduced set of variables, the errors of the ten best solutions for each sequence were again very low. Figure 9 shows the error of a representative solution (of the 50 runs) for each sequence after both experiments. The mean relative increase of the error from experiment 1 to 2 over all sequences is 50.3%. This may seem high, but the *absolute* increase of the error is only 0.1 mm. Visually, the match between the original and reproduced trajectories did not significantly degrade from experiment 1 to 2.

As in experiment 1, the optimized solutions of the 50 runs with different initial values for a particular sequence did not always converge. Figure 5 (c) illustrates this variation by means of the optimized target signals for the sequence [ˌnɔnɔˈnɔnɔ]. Here again, the major variations regard the onset times of the 3rd and 4th commands. However, in general, the variations were smaller than in the first experiment.

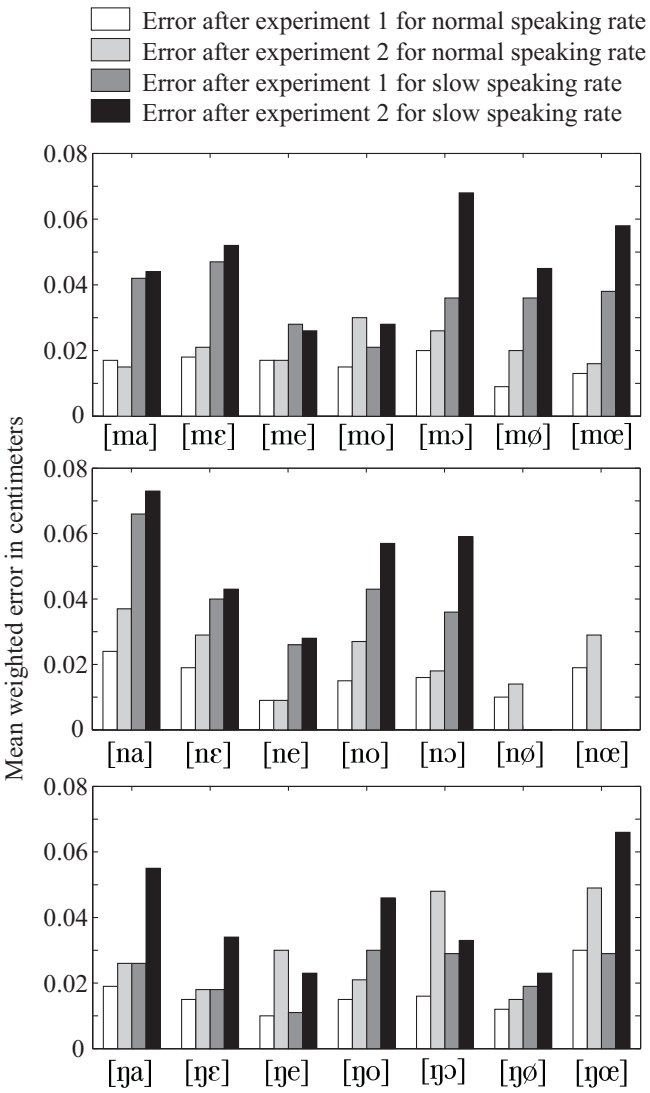


Fig. 9. Errors of the representative optimization results of each sequence after the two experiments. The mean error after experiment 1 over all sequences is 0.024 cm and 0.034 cm after experiment 2. The mean relative increase of the error from experiment 1 to 2 over all sequences is 50.3%, and the mean absolute increase of the error is 0.1 mm.

Figure 10 shows results for the sequences [ˌmɔmɔˈmɔmɔ], [ˌnɔnɔˈnɔnɔ], and [ˌŋɔŋɔˈŋɔŋɔ] for both normal and slow speaking rate. As presumed for the second experiment, the targets for the constrictor and jaw are equal for the normal and slow speaking rate variants of each consonant-vowel combination. Note also that the commands for the jaw and the constrictor are synchronized in each sequence as required. In most cases, the targets are not achieved by the articulators, especially for the normal speaking rate.

As opposed to the model assumptions, the *real* movements of the tongue tip, the lower lip, and the

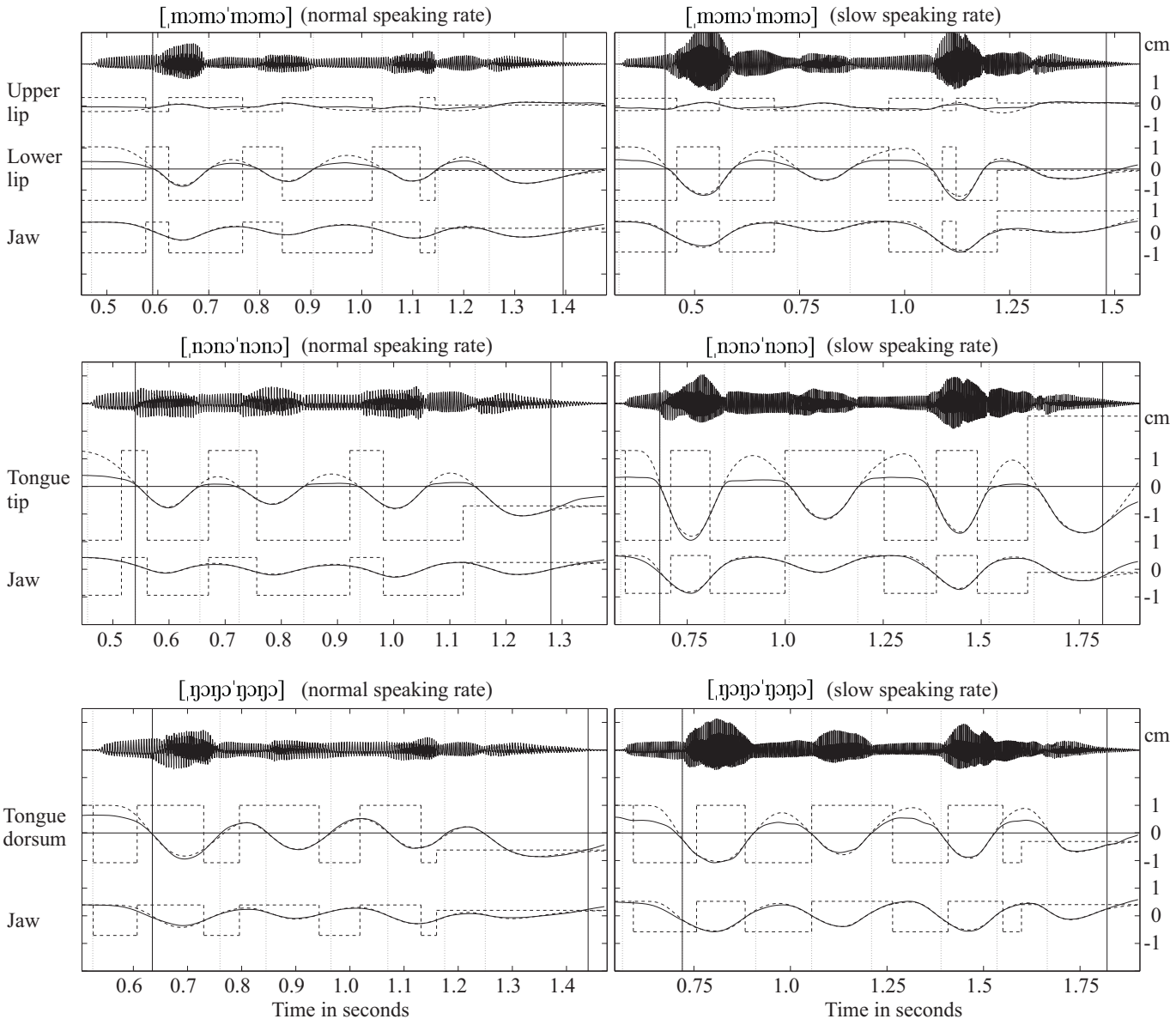


Fig. 10. Results for the sequences [ˌmɔmɔˈmɔmɔ], [ˌnɔnɔˈnɔnɔ], and [ˌŋɔŋɔˈŋɔŋɔ] at normal and slow speaking rates after the second experiment. Recorded trajectories are displayed as solid curves and the model-based reproductions as dashed curves. The target signals are displayed as dashed rectangular functions.

tongue dorsum in Fig. 10 do not always stop during the consonantal closure intervals. Especially the tongue dorsum trajectory during [ˌŋɔŋɔˈŋɔŋɔ] has no distinct closure plateaus. This can be attributed to the positions of the EMA coils on the tongue and the lips relative to the tongue and lip points that initiate the closures. When the positions of the observed fleshpoints differ slightly from the fleshpoints that initiate

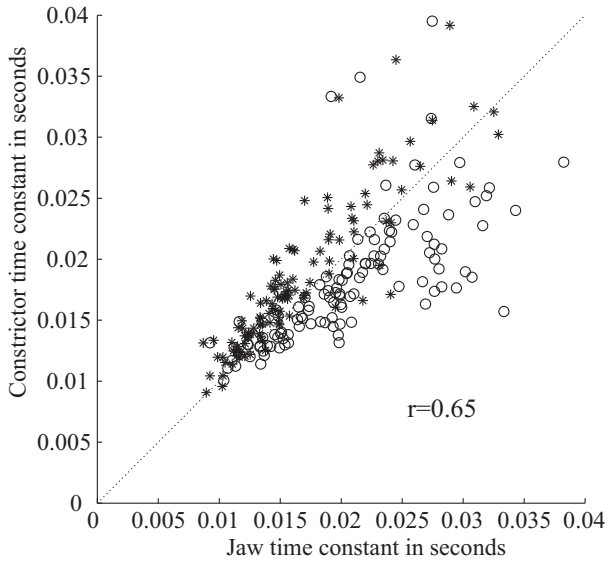


Fig. 11. Results for time constants after experiment 2. For all phones in all sequences, the time constant of the constrictor command is plotted against the time constant of the corresponding jaw command. Time constants of commands for opening gestures are displayed as circles and for closing gestures as stars.

the closures, the former continue to move somewhat after closure onset. Figure 10 indicates that this positional difference also depends on the speaking rate. In essence, the recorded curves show rather the movement of points *near* the fleshpoints that initiate the closures, while the clipping stage of the model refers to trajectories of the latter points. To model the observed effect, it would be conceivable to use, instead of clipping, some sort of “compression” of the model trajectories during the closure intervals that depends on the precise model fleshpoint positions.

Figure 11 plots the time constants of the jaw commands against the time constants of the constrictor commands. The correlation factor of 0.65 is somewhat less than in experiment 1, but the jaw and constrictor movements can still be regarded as strongly coupled. However, the principal difference between commands for opening and closing gestures observed already in experiment 1 is even more pronounced here. Regarding commands for opening gestures, the time constants for the jaw are greater than those for the constrictors in 84.2% of the cases. Hence, the jaw mostly opens somewhat slower than the constrictor. On the other hand, in commands for closing gestures, the time constants for the jaw are *smaller* than those for the constrictors in 82.5% of the cases. Therefore, the jaw mostly closes faster than the constrictor.

IV. Discussion and conclusions

This article introduces a dynamical model for the reproduction of articulatory trajectories in repeated consonant-vowel syllables based on sequential target approximation, which has the following essential properties:

- Tenth-order linear systems are used to model the dynamics of articulators. In comparison to traditionally used second-order systems, they are better qualified to reproduce the bell-shaped velocity profiles of natural directed movements for step functions as input. Second-order systems would require more complicated command specifications to achieve the same performance as tenth-order systems (e.g. [6], [11]).
- The model can reproduce the typical discontinuities in the trajectories of constrictors at the onset and offset of closures by a clipping mechanism. Together with the concept of virtual targets, the mechanism proved to be very effective to model both smooth and abrupt observed articulatory transitions in a uniform framework. In other attempts to model observed articulatory movements, this is usually entirely neglected [1], [2], [20], [21].
- The model transfers the system state between adjacent articulatory commands. This is physically more plausible than models based on superposition of impulse responses to generate articulatory trajectories.

In combination, these properties allow the detailed reproduction of movements with less degrees of freedom in the movement specification than previous models. With the constraints imposed in the second experiment, only three variables were necessary to capture the (prosodic) variations in the trajectories of the jaw and the constrictor between two phones, namely the common onset time and the time constants of the articulatory commands for both articulators.

With regard to the articulatory commands, we tested the assumptions that phoneme targets are invariant with respect to stress and speaking rate, that constrictor targets are furthermore invariant to vocalic context, and that articulatory commands for the constrictor and the jaw start synchronously. Our results indicate that these assumptions do not prevent the detailed reproduction of the observed speech movements. However, due to the limited corpus used in this study, these findings may not be overvalued. Note also that the invariance of the targets was defined along one-dimensional movement vectors in the midsagittal plane. In two-dimensional space, these targets are small regions rather than points, which is in agreement with previous findings (e.g. [17]). Furthermore, due to the biological nature of the human speech production system and the one-to-many mapping in the acoustic-articulatory relationship, the targets must be assumed

to have a small random component. The fact that a detailed fit of the trajectories could be achieved irrespective of random variations can be explained by the property of the model, that small variations of target values can be compensated by small variations of the time constants.

This study is the first step in the development of an improved quantitative control model for our articulatory speech synthesizer [35]–[37]. The generation of articulatory trajectories based on specifications of articulatory commands presented here is the lowest level of control in this scenario. For text-to-speech synthesis, higher levels of control are necessary. At a higher level, one would for example only specify the phone sequence, the speaking rate, and the intonation pattern and use a suitable machine-learning technique to predict the corresponding low-level command parameters. At this stage, it will be important to also model random variations of command parameters, as this variability is a key aspect of speech production [12]–[14]. At the control level discussed in this article, additional work is needed to complete and validate the model with respect to other classes of speech sounds like plosives, fricatives, and laterals. Also the examination of articulatory commands in consonant clusters is an important open issue.

Besides its benefit for speech production, the proposed model might also be useful in the field of speech recognition, where a growing interest for speech production models emerges [13]–[15]. It was recognized that speech production knowledge in automatic speech recognition may alleviate some common problems of current mainstream HMM-based speech recognizers and enable improved recognition of spontaneous speech and greater robustness to noise.

When the proposed model is interpreted in the context of motor control theory, the sequences of commands can be regarded as a motor program. According to Kandel et al. [38, p. 659], a motor program specifies the spatial features of a movement and the forces required to produce the desired movements. With respect to our model, the spatial features correspond to targets and the forces are related to the time constants. In this light, our results reveal an interesting difference of the forces acting on the constrictor and the jaw in opening and closing gestures, as depicted in Fig. 11.

ACKNOWLEDGMENT

The authors would like to thank Sascha Fagel at TU Berlin for providing the EMA data for this study, and the anonymous reviewers for many constructive suggestions. This work was supported in part by the German Research Council DFG Grant Nr. KR 1439/15-1.

REFERENCES

[1] T. Okadome and M. Honda, “Generation of articulatory movements by using a kinematic triphone model,” *Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 453–463, 2001.

[2] C. S. Blackburn and S. Young, "A self-learning predictive model of articulator movements during speech production," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1659–1670, 2000.

[3] Y. Xu, "Speech as articulatory encoding of communicative functions," in *The 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007, pp. 25–30.

[4] L. J. Lee, P. Fieguth, and L. Deng, "A functional articulatory dynamic model for speech production," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, 2001, pp. 797–800.

[5] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[6] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1878–1889, 1995.

[7] J. Simko and F. Cummins, "Sequencing of articulatory gestures using cost optimization," in *Interspeech 2009*, Brighton, United Kingdom, 2009, pp. 60–63.

[8] 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.

[9] A. G. Feldman, "Once more on the equilibrium-point hypothesis for motor control," *Journal of Motor Behaviour*, vol. 18, pp. 17–54, 1986.

[10] P. Perrier, D. J. Ostry, and R. Laboissière, "The equilibrium point hypothesis ond its application to speech motor control," *Journal of Speech and Hearing Research*, vol. 39, pp. 365–378, 1996.

[11] P. Perrier, H. Lœvenbruck, and Y. Payan, "Control of tongue movements in speech: The equilibrium point hypothesis perspective," *Journal of Phonetics*, vol. 24, pp. 53–75, 1996.

[12] G. Ramsay and L. Deng, "Tracking nonstationary targets using a dynamical system with markov-modulated parameters," *IEEE Signal Processing Letters*, vol. 2, no. 9, pp. 172–175, 1995.

[13] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 22, pp. 93–111, 1997.

[14] L. Deng, D. Yu, and A. Acero, "A bidirectional target-filtering model of speech coarticulation and reduction: Two-stage implementation for phonetic recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 256–265, 2006.

[15] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[16] B. Lindblom, "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[17] J. S. Perkell, M. L. Matthies, M. A. Svirsky, and M. I. Jordan, "Goal-based speech motor control: a theoretical framework and some preliminary data," *Journal of Phonetics*, vol. 23, pp. 23–35, 1995.

[18] J. Wei, X. Lu, and J. Dang, "A model-based learning process for modeling coarticulation of human speech," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 10, pp. 1582–1591, 2007.

[19] J. Dang, J. Wei, T. Suzuki, and P. Perrier, "Investigation and modeling of coarticulation during speech," in *Interspeech 2005*, Lisbon,Portugal, 2005, pp. 1025–1028.

[20] S. Bouabana and S. Maeda, "Multi-pulse LPC modeling of articulatory movements," *Speech Communication*, vol. 24, pp. 227–248, 1998.

[21] K. Ogata and Y. Sonoda, "Evaluation of articulatory dynamics and timing based on cascaded first-order systems," in

Proceedings of the 5th Seminar on Speech Production: Models and Data and CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, Kloster Seeon, Germany, 2000, pp. 321–324.

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

[23] A. Löfqvist and V. L. Gracco, "Interarticulator programming in VCV sequences: Lip and tongue movements," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1864–1876, 1999.

[24] ——, "Control of oral closure in lingual stop consonant production," *Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2811–2827, 2002.

[25] S. Fuchs, P. Perrier, and C. Mooshammer, "The role of the palate in tongue kinematics: An experimental assessment in VC sequences," in *Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 1487–1490.

[26] G. Heike, "Sprachsynthese auf der Basis eines Artikulationsmodells," in *Fortschritte der Akustik, DAGA '78*, Bochum, Germany, 1978, pp. 467–470.

[27] K. N. Stevens, *Acoustic Phonetics*. The MIT-Press, Cambridge, Massachusetts, 1998.

[28] 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.

[29] J. H. Milsum, *Biological Control Systems Analysis*. McGraw-Hill, 1970.

[30] P. J. Alfonso and T. Baer, "Dynamics of vowel articulation," *Language and Speech*, vol. 25, no. 2, pp. 151–173, 1982.

[31] R. Netsell and B. Daniel, "Neural and mechanical response time for speech production," *Journal of Speech and Hearing Research*, vol. 17, pp. 608–618, 1974.

[32] S. Fagel, *Audiovisuelle Sprachsynthese*. Logos Verlag Berlin, 2004.

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

[34] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

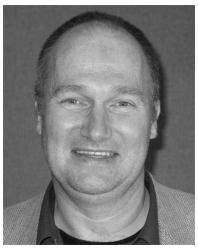
[35] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 2006, pp. 873–876.

[36] P. Birkholz and B. J. Kröger, "Vocal tract model adaptation using magnetic resonance imaging," in *7th International Seminar on Speech Production (ISSP'06)*, Ubatuba, Brazil, 2006, pp. 493–500.

[37] P. Birkholz, D. Jackèl, and B. J. Kröger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1218–1226, 2007.

[38] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science*. McGraw-Hill, 2000.

Peter Birkholz received his M.S. degree in computer science in 2002 and the Ph.D. degree (with distinction) in signal processing in 2005 from the Institute for Computer Science, University of Rostock, Germany. He has been working as research associate at the University of Rostock since 2005. Since 2009, he is with the Department of Phoniatrics, Pedaudiology, and Comunication Disorders, RWTH Aachen University, Germany. His main topics of reseach include articulatory speech synthesis and computational neuroscience. For his dissertation on articulatory speech synthesis, Peter Birkholz was awarded the Joachim-Jungius Prize 2006 by the University of Rostock and the Klaus-Tschira Award for Achievements in Public Understanding of Science in 2006.



Bernd J. Kröger received his M.S. degree in physics from the Rheinische-Wilhelms-University of Münster, Germany in 1985. He received his Ph.D. degree and his postdoctoral lecture qualification (Habilitation) in phonetics from the University of Cologne, Germany, in 1989 and 1998. Since 1992 he has been with the Department of Phonetics, University of Cologne, as an assistant professor and since 2001 he is with the Department of Phoniatrics, Pedaudiology, and Comunication Disorders, RWTH Aachen University, Germany, as a senior researcher and associate professor. His research interests are in the field of general phonetics, articulatory speech synthesis, and neural network modeling.

Christiane Neuschaefer-Rube (MD) studied medicine at the Friedrich-Wilhelms-University of Bonn, Germany from 1979-1986, granted by the Evangelische Studienwerk Villigst. In 1987 she became a Doctor of Medicine at the same university. Between 1987 and 1991 she specialized into ENT-medicine and between 1992 and 1995 into phoniatrics and pedaudiology. In 2000 she received her postdoctoral lecture qualification (Habilitation) in phoniatrics and pedaudiology at the RWTH Aachen University. Since 2001 she is full professor and head of the Clinic of Phoniatrics, Pedaudiology and Communication Disorders of the University Hospital Aachen and lecturer in the study courses medicine and bachelor and master of logopedics. She belongs to the examiner board of the medical specialists' council of North Rhine for medical doctors becoming specialists of phoniatrics and pedaudiology. Furthermore she is a member of the Board of the German association of Phoniatrics and Pedaudiologists (DGPP), and member of the EACCME-Board of the U.E.M.S. ORL section. Her main scientific interests are vocal tract function, gender vocology, voice perception, and articulatory modeling of voice and speech.