A gesture-based dynamic model describing articulatory movement data

Bernd J. Kröger, Georg Schröder, and Claudia Opgen-Rhein
Institut für Phonetik, Universität zu Köln, Greinstrasse 2, D-50939 Köln, Germany

(Received 3 May 1994; accepted for publication 13 April 1995)

A quantitative dynamic model for the description of speech movements using a critically damped linear second-order system is proposed. This six-parameter model is able to fit natural movement data with high accuracy. Since in this approach the actual location of gestural onset and gestural offset, i.e., the location and duration of gestural activation, results from the fitting procedure, no advance sectioning of movement traces is necessary. The model parameters are target position, eigenperiod, and four time parameters describing the temporal location of gestural onset and offset. The fitting algorithm is tested on simulated and natural data in order to evaluate the accuracy of the fits and the repeatability of the dynamic parameters extracted. © *1995 Acoustical Society of America.*

PACS numbers: 43.70.Bk, 43.70.Aj, 43.70.Hs

INTRODUCTION

A. Background

A quantitative dynamic model for articulation aims to describe the spatial and temporal properties of articulatory movements by means of a few, effective parameters. While kinematics describes movement merely as a function of time, dynamics deals with the forces within a system which cause such movement. Dynamics predicts movement by solving a parametrized equation of motion. For the mass–spring system, the dynamic parameters, mass, stiffness, damping, and rest position, are time-invariant. These four parameters together with the boundary conditions (e.g., initial displacement and velocity) are sufficient to exactly predict the movement behavior of the system.

A purely physical model is inadequate in the case of speech movements. Since speech movements are directly related to linguistic intentions, the dynamic parameters should be functions of linguistic parameters (e.g., phoneme string, stress, position). The gestural framework provides a useful basis for connecting a dynamic model and a linguistic description of speech articulation (Browman and Goldstein, 1989; Saltzman and Munhall, 1989; Kelso *et al.*, 1986). Within this framework, gestures are considered the basic units of speech, and a dynamic concept for articulation is proposed. Each gesture is modeled by a second-order dynamic system. Dynamic parameters, i.e., target (rest position), stiffness, and phase values describing intragestural temporal extent and intergestural timing, are defined for each gesture.

The approach for modeling articulator movements introduced in this paper differs from others (e.g., Sonoda, 1977; Shigenaga and Ariizumi, 1977; Flanagan *et al.*, 1990; Perrier *et al.*, 1991; Vatikiotis-Bateson *et al.*, 1991; Shirai, 1993) by a serial piecewise fitting of discrete sections (i.e., time windows). This sectioning guarantees an unambiguous relationship between a continuous physical dynamic and a discrete linguistic description. The gestural approach thus leaves the domain of time-invariant systems. Since gestures are active only within finite time intervals, and since gestures relating to different linguistic categories are described by quantitatively different sets of dynamic parameters, the system parameters change when one gesture ends and the following one becomes active.

There are different ways to define time windows or sections for dynamic parameter estimation. Kelso *et al.* (1985) use displacement extrema to define the time intervals for opening ("peak-to-valley") and closing ("valley-to-peak") gestures. Browman and Goldstein (1985, p. 110) define two different kinds of gestural time intervals. Like Kelso *et al.* (1985) they change the model parameters at displacement peaks ("transition hypothesis"), but, additionally, they propose an alternative division of articulatory traces by using the velocity extrema ("C–V hypothesis"). Smith *et al.* (1993) propose two different methods for sectioning traces, of which one again uses displacement extrema ("peak windows"), while the second one includes "the relatively flat plateau regions around displacement extrema with the region of movement between these plateaus (CV windows)" (Smith *et al.*, 1993, p. 1581) where plateau regions are arbitrarily defined by the criterion, that plateaus begin or end within 1% of the range of amplitude after the extreme peak or valley of the movement trace.

B. Motivation for this study

In order to establish a quantitative gestural model capable of fitting a great variety of natural movement traces, and in order to extract relevant dynamic parameters from movement data, three problematic points must be considered. First, the sectioning of movement traces, i.e., the location of the gestural activation interval should be the result of a parameter extraction procedure rather than based on an *a priori* assumption, since this location is one of the most important features within the gestural approach. Second, a point location for gestural onset and offset is an untenable assumption. Gestural activation pulses should have nonabrupt onset and offset portions (Saltzman and Munhall, 1989, p. 343) and

 0001-4966/95/98(4)/1878/12/$6.00

gestures — even if acting on the same articulator — can overlap in time (i.e., they can be blended. Saltzman and Munhall, 1989, p. 345f). As a consequence, the parameters of the underlying dynamic model change continuously even during activation of one gesture. In the approach described in this paper, we introduce gestural onset and offset time intervals. Within these intervals gestural parameters increase and decrease continuously.

Third, our experiments indicated that articulatory traces cannot be fitted with high accuracy by exponential time functions, i.e., movements resulting from a time-*in*variant critically damped second-order dynamic system. According to the succession of opening and closing gestures, most articulatory traces show a more or less oscillatory pattern. They seem to be comparable to an undamped sinusoidal motion rather than to a critically damped target-directed movement. But the use of undercritical damping (Smith *et al.*, 1993) violates one of the basic ideas within a gestural theory: Each gesture corresponds to one monotonic (ascending or descending) movement pattern. The oscillatory shape of movement traces is caused by the succession of different gestures acting on the same articulator. It will be shown in this paper that the introduction of gestural onset and offset portions makes the movement shapes more flexible and allows modeling a great variety of natural movement traces without violating the assumption of monotonicity.

I. THE MODEL

A. The force field approach

1. Basic concepts of the gestural approach

Informally, a gesture is identified with the formation (and release) of a characteristic vocal tract constriction (e.g., a bilabial closure), where different combinations of articulators perform a gesture (e.g., jaw, lower and upper lips for a bilabial gesture). In this sense, the term gesture can be used to denote "a member of a family of functionally equivalent articulatory movement patterns that are actively controlled with reference to a given speech-relevant goal" (Saltzman and Munhall, 1989, p. 334). In general, the gestural concept includes neuronal, muscular, and dynamical aspects. Each gesture is originated by neuronal activity, which governs specific muscular actions causing goal-directed articulator movements. But our quantitative description of gestures mainly concentrates on the latter aspect. the dynamics of articulator movements. In this preliminary stage we avoid concrete physiological modeling. The temporal extent of gestures is reflected by *gestural activation intervals*, denoting the time interval in which each gesture is active (Browman and Goldstein, 1990; Salzman and Munhall, 1989). In order to model the temporal variation of activity during each gestural activation interval, we introduce a *gestural strength function*, which characterizes the time-varying strength of an underlying gestural force field quantitatively.

Figure 1 introduces the basic concepts of the gestural approach in the context of gestural resynthesis. The upper half of the figure gives the vertical component of movement traces of tongue tip and tongue body for an utterance. Below, control parameter time functions for the articulatory resynthesis of this utterance are shown, using temporal and dynamical information extracted from the data for the control parameters tongue height and tongue tip height. Here a very simple quantitative gestural model is used. Gestural onset and gestural offset are discrete time instants and the gestural force field is constant during the whole activation interval (i.e., the assumption of a point location for gestural onset and offset; step-rectangular-shaped gestural strength functions). The shaded boxes indicate the temporal extent of each gesture, i.e., the location of gestural activation intervals. Gestural onset and offset are determined by the criterion of extreme displacement of the vertical component of the movement traces.

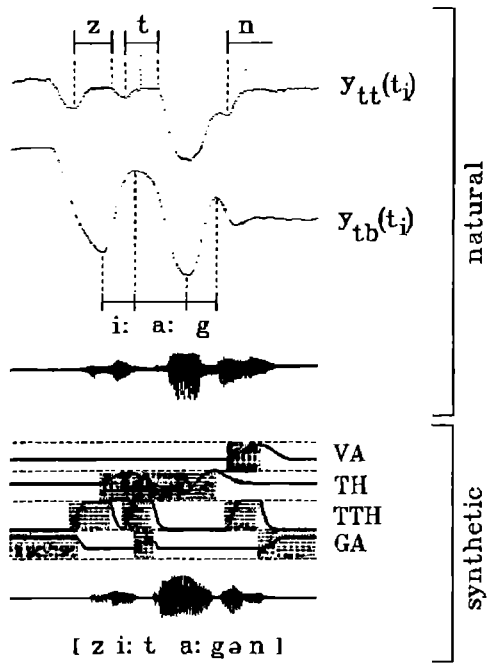

FIG. 1. Articulatory resynthesis for the German utterance /zi:ta:gən/ using step-rectangular-shaped gestural force functions. Top: naturally produced movement traces for tongue tip $y_{tt}(t_i)$ and tongue body $y_{tb}(t_i)$ (t_i indicates sampling time instants). Horizontal bars: activation intervals for three apical gestures (for production of /z/, /t/, and /n/), for three dorsal gestures (for production of /i:/, /a/, and /g/). Bottom: control parameter time functions produced by the gestural control module of an articulatory synthesizer (Kröger, 1993a, b) for velic aperture (VA), tongue height (TH), tongue tip height (TTH), and glottal aperture (GA). Shaded areas: gestural activation intervals.

2. A quantitative dynamic model for gestures

Each gesture is defined by the parameters of its underlying dynamic model, i.e., the critically damped second-order system (for equations see Browman and Goldstein, 1990, p. 372, Kröger, 1993a, p. 232). We avoid the mass–spring analogy — and, consequently, the parameters mass and stiffness — since it has drawbacks in connection with speech gestures. First, stiffness describes the physiological and dynamical behavior of muscles (e.g., Flanagan *et al.*, 1990, p. 36f) and thus, together with mass, the low-level dynamic characteristics of articulators. But gestures represent a (theoretical) concept for the active high-level control of articulators. Therefore, the parameter "mass" is not needed explicitly within the quantitative formulation of ges-

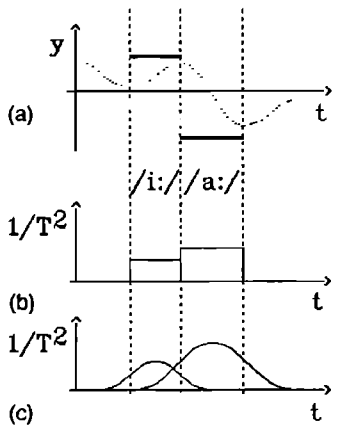

FIG. 2. Vertical component of tongue body movement and force functions for successive /i:/ and /a:/ gestures as a function of time. (a) Abscissa: time t; ordinate: displacement y; dotted line: vertical component of tongue body movement; thick horizontal bars: location of gestural targets. (b) and (c) Discontinuous and continuous force functions (i.e., strength of gestural force field as a function of time) for successive /i:/ and /a:/ gestures. The vertical dashed lines mark the gestural activation intervals.

tural dynamics given by Browman and Goldstein (1990, p. 372). Second, the physical mass–spring model implicates time invariance, whereas the dynamic model for gestures (gestural second-order system) is strongly time variant, since gestures are active only within definite intervals. In the case of critical damping the equation of motion can be written as

$$\ddot{y}+2\omega\dot{y}+\omega^2(y-y_{tg})=0, \quad (1)$$

with instantaneous displacement, velocity, acceleration of the mass y, $\dot{y}$, $\ddot{y}$, rest position (equilibrium position) of the system y_{tg}, and the eigenfrequency (Hz) of the undamped system $\omega/2\pi$. y is a one-dimensional variable and denotes the main movement direction of the articulator during gestural activation. Eigenfrequency can be replaced by eigenperiod T ($T=2\pi/\omega$). The strength of the force field per unit mass is given by ω^2 in Eq. (2) and thus equals $4\pi^2/T^2$. Consequently, the strength of the gestural force field per unit mass (SFF) is proportional to eigenfrequency squared or to reciprocal eigenperiod squared:

$$\mathrm{SFF}\sim\omega^2\sim(1/T)^2. \quad (2)$$

3. Time-varying force fields as a model for gestures: The problem of gestural onset

In general, articulatory movements exhibit oscillatory rather than monotonically ascending or descending patterns (e.g., tongue body movement in Fig. 1). This oscillation results from the temporal succession of gestures with different target positions (e.g., for tongue body the /i:/-gesture and /a:/-gesture in Fig. 1). The oscillation can be seen as a change of movement direction according to successively activated gestures with different target positions. By associating gestures with (abstract) time-varying force fields, the oscillatory movement pattern can be modeled. Figure 2 illustrates that the gesture-performing articulator is accelerated toward the instantaneous gestural target by the instantaneously active gestural force field. The oscillatory behavior of the articulator movement originates from the temporal

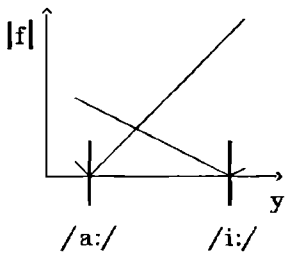

FIG. 3. Force fields for /i:/ and /a:/ gestures. Abscissa: vertical component of tongue body displacement [equals ordinate in Fig. 2(a)]; Ordinate: unsigned force. Vertical thick bars: location of gestural target [equal horizontal thick bars in Fig. 2(a)].

change from the /i:/-gesture to the /a:/-gesture force field. Two possible gestural strength functions are given for the /i:/- and /a:/-gesture, i.e., a step-rectangular-shaped constant gestural strength function and a sine-wave-shaped continuously time-varying strength function [Fig. 2(b) and (c)]. The gestural target (i.e., gestural rest position) is defined by the zero value of the appropriate gestural force field (Fig. 3).

For step-rectangular-shaped strength functions (discontinuous case, constant force field during gestural activation interval), two problems occur. First, the time function for articulator displacement is an ordinary exponential time function, thus limiting the range of possible shapes for gestural movement patterns. Such a model is too unflexible to fit natural articulatory data satisfactorily. Second, the force field changes abruptly, leading to unrealistically high acceleration of the gesture-performing articulator at gestural onset.

Figure 4 shows an attempt at fitting the /i:/- and /a/-gesture of an utterance by using step-rectangular-shaped strength functions. It can be seen that in both cases the fitting is unsatisfactory for displacement, velocity, and acceleration. That the model time functions are not satisfactory for velocity and acceleration can be understood from the fact that the fitting algorithm only tries to minimize the differences between model and data for *displacement*. But according to the dynamic character of the underlying model — the equation of motion includes displacement, velocity, and acceleration—the fit of acceleration as well as velocity should be better. Considering acceleration, it can be seen from Fig. 4 that especially at gestural onset, an undesirably pronounced acceleration-peak occurs. Even if the time window is chosen in such a way that gestural onset coincides with an acceleration peak of the data (see the /i:/-gesture), the acceleration peak produced by the model is too strong.

This peak is caused by the abruptness of gestural onset in this approach (see shaded areas in Fig. 4). Figure 5 illustrates the abrupt change of force, acting on the articulator in this case (thick dashed lines). This discontinuous change of forces can be avoided if the gestural onset time *instant* is replaced by a onset time *interval* in which the force field increases continuously from zero to its full value as illustrated by the solid thick line in Fig. 5(b), where the force field increases during the time interval marked by t_1 and t_3. This time interval is called the *gestural onset interval*. (Only because of its explanatory power the continuously increasing strength function is modeled here by using five discrete steps.)

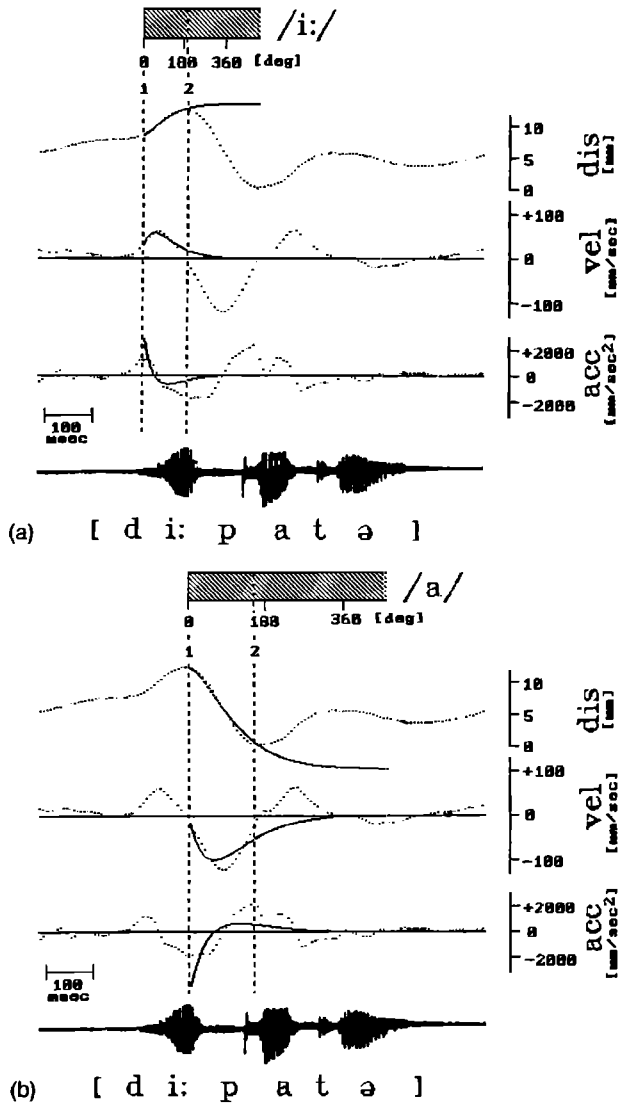

FIG. 4. (a) and (b) Fit of /i:/ and /a/ gesture by using step rectangular force functions for /di:patə/. Displacement, velocity, and acceleration as function of time for a naturally produced tongue body movement trace are indicated by dotted lines. Fitted model time functions are indicated by solid lines. Shaded areas indicate gestural force functions. Gestural phase values are given below the activation pulse shapes. Vertical dashed lines (markings) indicate gestural onset (marking 1) and gestural offset (marking 2).

It will be shown below that much better fits of these /i:/- and /a/-gestures are possible if a step-rectangular-shaped strength function is avoided and if a continuously shaped one is introduced as realized by the six-parameter model.

B. The six-parameter model and the fitting procedures

1. The shape of the strength function

Figure 6 shows the strength function in units of reciprocal eigenperiod squared T^{-2}, which is proportional to the instantaneous SFF [Eq. (2)]. The shape of the gestural strength function is defined by peak eigenperiod T_0, two time values defining the temporal location of the onset interval t_{1on} and t_{2on}, and two time values defining the temporal location of the offset interval t_{1off} and t_{2off}. The peak eigenperiod value T_0 gives the maximal strength of the force field. Three time portions occur within the gestural activation interval. Portion 1 is the onset interval, and portion 2 (not

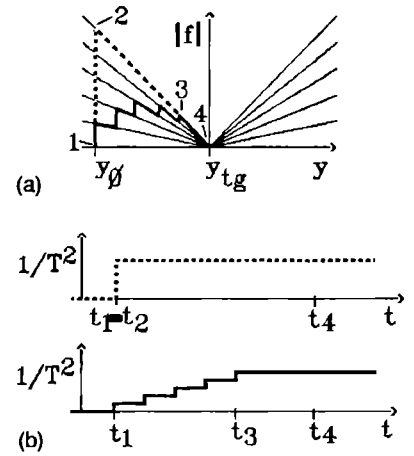

FIG. 5. (a) Five force fields (unsigned force f as function of displacement y; thin solid lines) with different strength. y_0 indicates initial displacement at beginning of gestural onset. y_{tg} indicates gestural target. The labels 1 to 4 mark time instants [see also (b)] and the thick lines mark traces through the force-displacement space for abrupt gestural onset (thick dashed line) and for continuously increasing gestural onset (thick solid line). (b) Gestural force functions. Thick dashed line: Step-rectangular force function, i.e., abrupt gestural onset; thick solid line: continuous force function, i.e., continuously increasing onset portion (approximated by five steps).

labeled in Fig. 6) is the steady-state interval in which gestural activation is at its maximum. The beginning and end of this portion is given by t_{2on} and t_{1off}. Portion 3 indicates the offset interval. Eigenfrequency ω is modeled by sine quarter waves within the onset and offset intervals. Equation (3) gives the time function of eigenfrequency for a gestural activation pulse:

$$\omega(t)=\omega_0 \sin\left(\frac{2\pi(t-t_{1on})}{4(t_{2on}-t_{1on})}\right), \quad \text{for } t_{1on}\leq t<t_{2on},$$

$$\omega(t)=\omega_0, \quad \text{for } t_{2on}\leq t<t_{1off}, \tag{3}$$

$$\omega(t)=\omega_0 \sin\left(\frac{2\pi(t-t_{2off})}{4(t_{1off}-t_{2off})}\right), \quad \text{for } t_{1off}\leq t<t_{2off}.$$

Since the instantaneous SFF is represented by ω^2, the gestural strength function is given by squared sine quarter waves which leads to a smooth change of the force field strength during the onset and offset interval. (Another simple model function is to use a cosine half-wave directly for SFF. But that leads to a square root function for eigenfrequency and to a numerically more complex phase value calculation.)

Figure 7 show fits of the /i:/- and /a/-gesture, now using the six-parameter model. Labels 1 and 2 indicate the temporal extent of the gestural onset interval. In contrast to the fits

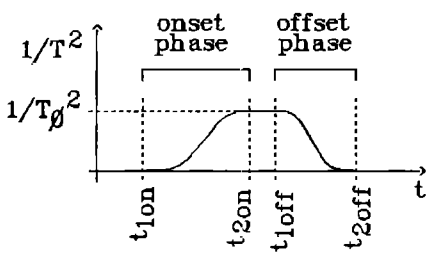

FIG. 6. Shape of gestural force function for the six-parameter model.

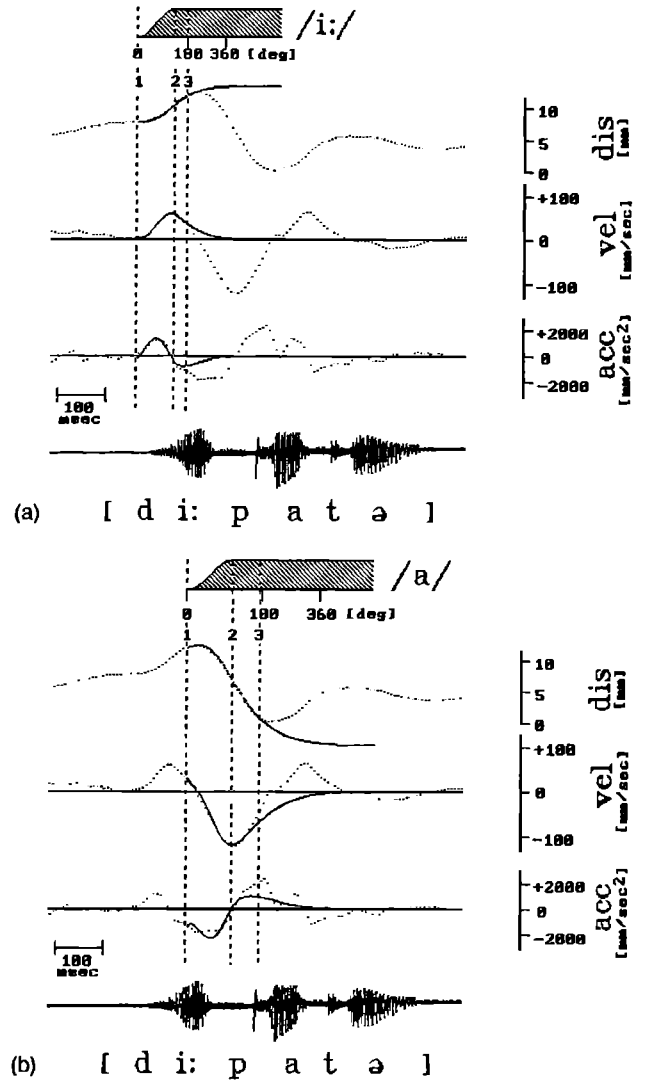

FIG. 7. (a) and (b) Fit of /i:/ and /a/ gesture by using a continuous force function (compare with Fig. 4). Marking 1 (2) indicates beginning (end) of gestural onset portion (t_{1on},t_{2on}); marking 3 indicates end of fit interval.

using step-rectangular-shaped strength functions (Fig. 4), now displacement, velocity, and acceleration are fitted very well, indicating that the continuous strength function is more realistic. The onset interval physically represents the acceleration interval during which the gesture-performing articulator changes its movement direction toward the target of this gesture and accelerates toward gestural peak velocity. Figure 7 also indicates that the gestural onset interval surrounds an acceleration peak. Since the /i:/-gesture is the first active tongue body gesture in the given utterance, the gestural onset interval equals the whole positive acceleration pulse. As a consequence of gestural overlap, the first part of the following negative acceleration pulse is modeled by the /i:/-gesture before the onset interval of the /a/-gesture starts.

Due to the time dependence of ω the solution of Eq. (1) cannot be written as a single analytical expression. Equation (1) is solved here numerically by discretizing its time derivatives using backward differences. This method is also applicable in the case of overlap (two instantaneously active force fields). In this case, a linear superposition of both force fields is assumed and the equation of motion is

$$\ddot{y}+2(\omega_1+\omega_2)\dot{y}+\omega_1^2(y-y_{tg1})+\omega_2^2(y-y_{tg2})=0, \quad (4)$$

with ω_i instantaneous (time-dependent) eigenfrequency and y_{tgi} target of the appropriate gesture ($i=1,2$).

2. The task of the fitting algorithm

The task of the fitting algorithm is (1) the estimation of the gestural target; and (2) the estimation of the temporal location and strength of the gestural activation. The latter implies (2a) the estimation of (peak) eigenperiod which corresponds to the estimation of the maximal strength of the gestural force field, and (2b) the estimation of four time values locating the gestural onset and offset intervals. Targets are assumed to be gesture-inherent parameters and therefore constant for each type of gesture (e.g., dorsal closing versus apical closing gesture). Therefore targets are estimated exclusively from unreduced tokens of a gesture, i.e., from gestures which occur within the most stressed syllable of an utterance.

3. Error criterion

Fitting is performed by minimizing the difference between model and data curve within the fitting interval (signal window):

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i-y(t_i,a_1,\ldots,a_M)}{s_i}\right)^2\rightarrow\min. \quad (5)$$

The signal window consists of N data points (t_i,y_i) $(i=1,\ldots,N)$ and the model consists of M adjustable parameters a_j, $j=1,\ldots,M$ (in this case $M=6$: y_{tg}, T_0, t_{1on}, t_{2on}, t_{1off}, and t_{2off}) two or more of which are held constant according to the actual fitting procedure. $y(t)=y(t,a_1\ldots a_M)$ denotes the model curve and s_i denotes the standard deviation for the data points (here s_i=const. for $i=1,\ldots,N$). Since onset location (t_{1on}, t_{2on}) and offset location (t_{1off}, t_{2off}) are adjustable and since N represents the window length (in sample points), depending only on onset and offset location, the value N changes during the fitting procedure. The term $1/N$ in Eq. (5) normalizes the data to model difference with respect to the instantaneous window length. No conventional fitting algorithm can be used, first, since the signal window length and location is not fixed, and, second, since the model function cannot be written as a single analytical expression. For these two reasons a multidimensional minimization algorithm, the downhill simplex method (Press *et al.*, 1992, p. 408ff), is used. Since the algorithm does not necessarily detect the global minimum, i.e., does not necessarily give the best fit parameters, we rejected fits for which the time parameters t_{1on}, t_{2on}, t_{1off}, and t_{2off} do not occur at the proper places, i.e., for which the resulting signal window is too small, or does not cover the gestural velocity peak. A proper location of gestural onset and offset is a strong criterion for the goodness of the fits. The fact that a signal window does not shrink during a fitting procedure, together with the fact that the signal window occurs at the proper place is a strong indicator that the fit and the derived model parameters are acceptable.

The minimization criterion [Eq. (5)], i.e., the minimization of displacement, is sufficient only for eigenperiod, onset,

and offset estimation. In the case of target estimation the model does not fit the gestural velocity peak satisfactorily in many cases. Peak velocities produced by the model were often too small in comparison to the actual data. Therefore we changed the minimization criterion for this procedure by considering displacement *and* velocity. Here, in addition to the displacement differences, the velocity differences between model and data must be minimized in a 10-ms interval around the gestural velocity peak.

4. The fitting procedures

Since the simultaneous optimization of six parameters per fit is difficult and since targets must be estimated from unreduced tokens of each type of gesture, parameter estimation is organized in three procedures:

Fitting procedure I: Target estimation is performed by simultaneous calculation of three parameters: target position y_{tg}, eigenperiod T_0, and beginning of onset interval t_{1on}. The other three parameters (i.e., end of onset interval t_{2on}, beginning and end of offset interval t_{1off}, t_{2off}) are set to their initial values and remain fixed.

Fitting procedure II: Eigenperiod and onset estimation is performed by simultaneous calculation of T_0, t_{1on}, t_{2on}, and t_{1off} (Fig. 7). The target position y_{tg} must be known in advance. Here, t_{1off} indicates the end of the fit interval. In many cases this time instant marks the onset of the following gesture or the beginning of clipping (i.e., contact of articulator with a rigid vocal tract wall, see Kröger, 1993a).

Fitting procedure III: Offset estimation of the current gesture and onset- and eigenperiod estimation of the following gesture is performed by calculating t_{1off} and t_{2off} simultaneously with the parameters T_0, t_{2on}, and t_{1off} of the following gesture. End of offset of a gesture can only be calculated in connection with the calculation of onset- and eigenperiod of the following gesture since the procedure takes the overlap of two gestures into account (Fig. 8). Estimation of t_{1off} and t_{2off} implicates that the four other parameters of the gesture must be known in advance (procedures I and II). Additionally, the target of the following gesture must be known (procedure I for the following gesture). Two constraints are postulated in order to limit the degrees of freedom for this fitting procedure; t_{1on} of the following gesture equals t_{1off} of the preceding gesture, and t_{2off} of the preceding gesture must occur earlier than t_{2on} of the following gesture (Fig. 8).

While procedures I and II are isolated piecewise fittings, successive application of procedure III leads to fits of whole utterances. Beside offset interval estimation, the important result of fitting procedure III is the estimation of eigenperiod and onset interval of gestures, which now takes into account the influence of overlap with the preceding gesture.

5. Starting values

For a successful fitting and parameter estimation, the proper choice of starting values for fitting, i.e., a careful initial adjustment of the model parameters, is very important. In a first step the velocity extremum, which is the most easily detectable physical event corresponding to a gesture, is located (see the arrows labeled /i:/ and /a/ in Fig. 9). In a second step, as a rough approximation for the temporal extent of the gesture, the velocity zero crossings, which surround the gestural velocity peak, are determined. These zero crossings mark displacement extrema and are indicated for the /i:/- and /a/-gesture by the vertical dashed lines in Fig. 9. The target value is initially set to the extreme displacement value at the time of gestural offset. Eigenperiod is initially set to twice this time interval length, since gestural offset roughly corresponds to a phase value of 180 deg. In a third step, a first estimate for gestural onset and offset intervals is made. The time of the velocity extremum (i.e., acceleration zero crossing), which corresponds to the central gestural ve-

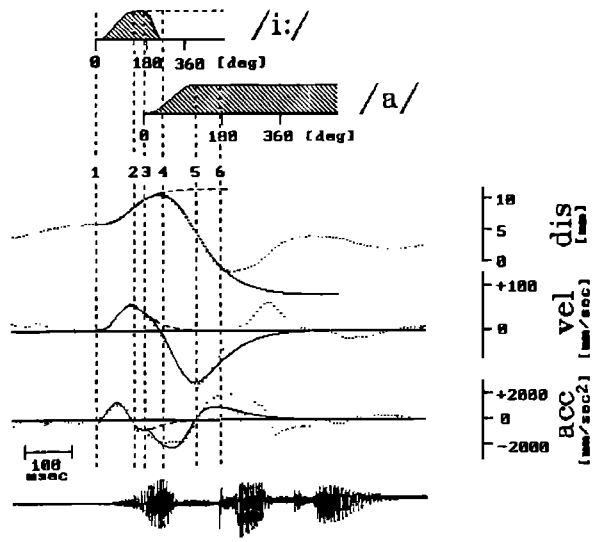

FIG. 8. Fit of /i:/ and /a/ gesture by using continuous force functions and by including gestural overlap. Marking 1 (2) indicates beginning (end) of onset of /i:/ gesture ($t_{1on/i:/}$,$t_{2on/i:/}$); marking 3 indicates beginning of offset of /i:/ gesture ($t_{1off/i:/}$) and equals beginning of onset of /a/ gesture ($t_{1on/a/}$); marking 4 indicates end of offset of /i:/ gesture ($t_{2off/i:/}$); marking 5 indicates end of onset of /a/ gesture ($t_{2on/a/}$); marking 6 indicates end of fit interval for /a/ gesture.

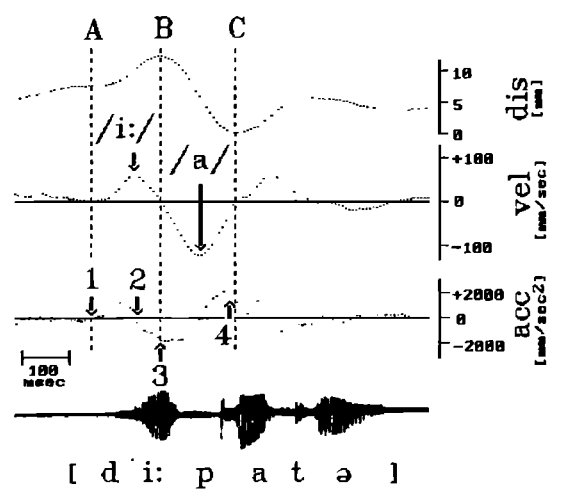

FIG. 9. Measurement data for fits given in Figs. 4, 7, and 8 (velocity and acceleration are calculated using cubic spline smoothing). The arrows mark gestural velocity peaks, acceleration peaks, and acceleration zero crossings, which are important for obtaining proper initial conditions. Markings A, B, and C indicate initial gestural time intervals.

locity extremum, is taken as initial time for t_{2on} (indicated by arrows labeled /i:/ and /a/ in Fig. 9). Then, the neighboring acceleration zero crossing, which precedes this time, is taken as initial time for t_{1on} (indicated by arrow 1 for /i:/ and by arrow 2 for /a/). The time of the acceleration extremum between the gestural velocity peak and the following velocity zero crossing, marks the end of initial fit interval for procedures I and II (indicated by arrow 3 for /i:/ and arrow 4 for /a/ gesture). In the case of procedure III, the end of the fit interval of the preceding gesture is taken as the initial time for t_{1off} of this (preceding) gesture, and the time of the velocity peak of the following gesture is taken as the initial time t_{2off} of the preceding gesture.

There are additional constraints, which must be met during the fitting procedure. First, the target value must not fall beyond the limit given by the extreme displacement value. Second, this temporal ordering of the four time parameters must remain undisturbed and a minimal temporal distance (dt) between these labels must be maintained:

$$t_{1on}+dt<t_{2on}; \quad t_{2on}+dt<t_{1off}; \quad t_{1off}+dt<t_{2off}, \quad (6)$$

with $dt=5$ ms. Third, T_0 must be neither zero nor negative.

6. Phase value calculation

In this quantitative approach, phase values can be interpreted as a measure for the relative articulator-target distance (relative to initial articulator-target distance) and therefore as a measure for the degree to which a gesture has been executed (Kröger, 1993a and 1993b). A phase scale can be calculated for each concrete gesture of an utterance. The location of gestural phase scales depends on the location of the gestural onset interval; its enlargement factor depends on eigenperiod. Phase values can be calculated for a gesture by

$$\varphi(t)=\int_0^t \omega(\tau)d\tau. \quad (7)$$

From Eq. (7) the phase as a function of time can be written as a single expression for our six-parameter model, using Eq. (3), the sine quarter wave case:

$$\varphi(t)=\frac{4(t_{2on}-t_{1on})}{T_0}\left[1-\cos\left(\frac{2\pi t}{4(t_{2on}-t_{1on})}\right)\right], \quad \text{for } t_{1on}<t\leq t_{2on}$$
$$\varphi(t)=\frac{4(t_{2on}-t_{1on})}{T_0}+\frac{2\pi(t-t_{2on})}{T_0}, \quad \text{for } t_{2on}<t. \quad (8)$$

The absolute duration (ms) for a constant phase value distance (e.g., 10 deg) decreases during the gestural onset time interval, which reflects the increasing strength of the gestural force field during gestural onset. Phase values increase slowly at the beginning of gestural onset and more rapidly at its end. Equation (8) shows that phase value calculation (i.e., the location of gestural phase scale in time) solely depends on peak eigenperiod T_0, and gestural onset location t_{1on} and t_{2on}. The offset interval is not important for the phase scale since the phase scale is intended to deliver a measure for the (theoretically) temporally continuing gesture (solid lines in Fig. 7).

TABLE I. Corpus I: Broad transcription of analyzed German words /'CV:Cən/ uttered within different carrier phrases, ordered for analyzed gestures (rows) and its context (columns). Analysis is performed for (a) three consonantal gestures (labial, apical, and dorsal) and (b) two vocalic gestures (dorsal /a:/ and dorsal /i:/ gesture), which form the stressed syllable /'CV:/ of the word, directly following the carrier phrase. Carrier phrases: /ʔIç,za:..../ ("I saw ..."); /di:..../ ("The ..."); /fɐ.../ (German prefix).

Gesture	Carrier phrase	Voiceless Ca:	Voiceless Ci:	Voiced Ca:	Voiced Ci:
(a) Consonantal					
labial	ʔIçza:...	'pa:tən	'pi:pən	'ba:nən	'bi:nən
apical	ʔIçza:...	'ta:tən	'ti:fən	'da:mən	'di:vən
dorsal	ʔIçza:...	'ka:mən	'ki:mən	'ga:bən	'gi:sən
(b) Vocalic					
/a:/	di:....	'pa:tən		'ba:zən	
/i:/	f*....		'pi:pən		'bi:tən

II. EXPERIMENTS

A. Material and procedure

In order to estimate gestural parameters from naturally produced speech movements, two data corpora were collected from three adult speakers (two males, BK and GS; and one female, CO, all native speakers of German with no known speech anomalies). Corpus I: The speakers produced the phrases given in Table Ia three times and the phrases given in Table Ib six times. The procedure was repeated for each speaker 1 week later (two sessions). We analyzed five types of gestures occurring within the most stressed syllable of the utterance (one gesture from each item, as described in Table I): three types of consonantal gestures, i.e., apical, dorsal, and labial (closing) gestures, and two types of vocalic gestures, i.e., dorsal /i:/ and dorsal /a:/ gestures. Corpus II: The same three speakers produced two short phrases 30 times: /di:'patə/ (slang form for "the money"), and /di:'pa:tən/ ("the godfathers"). This procedure was repeated 1 week later. Two types of vocalic gestures occurring within the most stressed syllable /'pa/ of "die Patte" and /'pa:/ of "die Paten", i.e., German long- and short-vowel gestures, were analyzed. In order to measure the accuracy of the estimated dynamic model parameters, a third data corpus, containing synthetic articulator movements, was established. Here, two overlapping gestures were generated by the six-parameter model. The range of parameter values used for the generated movements comprised the parameter range of the gestures analyzed from corpus I and corpus II.

Lip and tongue movements were tracked using an alternating magnetic field device, the Articulograph AG-100 (Carstens Medizinelektronik GmbH, Göttingen, Germany) (Tuller *et al.*, 1990; Schönle *et al.*, 1987; Perkell *et al.*, 1992, p. 3093f). Receiver coils (diameter around 2 mm) were placed on the lower lip, on the tongue tip, and two on the tongue dorsum (around 20–25 mm and 40–50 mm posterior to the tongue tip receiver coil location), the anterior of which was used for tracking tongue body movements. In addition,

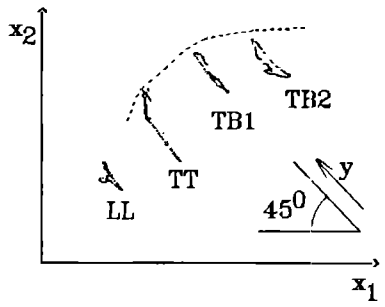

FIG. 10. Movement trajectories for lower lips (LL), tongue tip (TT), and tongue body (TB1 and TB2) for /di:pata/ (subject BK). The x_1/x_2-coordinate system is defined by the Articulograph helmet; y gives the main movement direction (the location of the palate is indicated by the dashed line).

one receiver coil was placed on the upper part of the nose for measuring a reference position. The receiver coil positions were tracked as function of time within the midsagittal plane. The magnetic field is produced by three transmitter coils mounted on a helmet worn by the speaker. For later computer analysis the movement data were directly AD converted (400-Hz sampling frequency). The acoustic signal was recorded simultaneously (16 bits, 16 kHz).

A direct interpretation of the two displacement-time functions provided by the device $[x_1(t).x_2(t)]$ is problematic since the x_1/x_2-coordinate system of the device is arbitrary. In the case of the Articulograph this coordinate system is defined solely by the helmet position. Thus the two-dimensional movement data were reduced by calculating the component in the main movement direction for each speaker and each session (y direction in Fig. 10), which normalizes the data according to helmet position and rotation. Since we focused on consonantal closing gestures in /i/–/a/ context and vocalic gestures forming /i/–/a/ or /a/–/i/ transitions, the main movement direction is clearly visible from the (two-dimensional) midsagittal trajectories for each receiver coil. We identified this main movement direction with the model control parameter dimensions lip aperture, tongue tip height, and tongue body height (Kröger, 1993a), i.e., with the dimension "degree of constriction."

The resulting one-dimensional movement data as a function of time were smoothed by natural cubic spline interpolation (see Press *et al.*, 1992, p. 113ff; with a temporal window of 25 ms). Fitting was done on the unsmoothed movement data. But smoothing was helpful in order to implement semiautomatic procedures for peak picking in order to calculate the proper starting values.

B. Results and discussion

1. Accuracy of fitted movement traces and dynamic parameters

The accuracy of the fitted movement traces was evaluated for the gestures of corpus I by calculating the mean distance of model to data values within the signal window using fitting procedure II (Table II). Comparing this mean distance to the mean distances of raw data to (cubic spline) smoothed data shows that the six-parameter model approximates natural movement traces with nearly the same accuracy as the smoothing algorithm itself. The mean difference between model and data is around 0.013 mm (±0.003 mm), while the mean difference obtained by cubic spline smoothing is only slightly lower (0.009 mm ±0.004 mm). These mean values and their standard deviations were calculated across the whole signal window, taking into account all subjects, sessions, and types of gestures. Analysis of different types of gestures showed that the mean values are somewhat greater for apical gestures, since these gestures exhibit strong changes in velocity and acceleration.

The accuracy of parameters estimated by procedures I (target position) and II (eigenperiod, onset interval location and beginning of offset interval) can be measured by fitting synthetic movement traces. For synthetic movements, the underlying dynamic parameter values are known in advance (predefined values) since these movements are generated by the six-parameter model. These values can be compared with the parameter values calculated by applying the fitting procedures (calculated values) to these synthetic time functions. Table III gives the mean difference of calculated to predefined parameter values diff_{syn} in absolute and relative units. The synthetic movements (corpus III) were generated by using values within the parameter range given in Table III. No range can be given for t_{1on}, since the absolute time value of this parameter serves as reference (is set to zero). For estimation of target position y_{tg} the mean absolute difference between calculated and predefined values can be related to the distance of target position to initial gestural displacement (given as range in Table III). That leads to a

TABLE II. Data and model values for all speakers, all sessions (corpus I, N=1080). Data values: gestural peak-to-peak amplitude dy_{pp}, gestural peak-to-peak interval length dt_{pp}; peak velocity v_{max}. Model values: fit-interval length t_{1off} -t_{1on}, mean distance of model to data values using cubic spline fitting dy_{csp}, and mean distance of model to data values using the six-parameter model dy_{6pm} within the fit interval. (ab: absolute values; s.d.: standard deviation.)

	dy_{pp} [mm]		dt_{pp} [ms]		v_{max} [mm/s]		t_{1off} - t_{1on} [ms]		dy_{csp} [μm]		dy_{pm} [μm]	
Gesture	ab	s.d.	ab	s.d.	ab	s.d.	ab	s.d.	ab	s.d.	ab	s.d.
Labial	5.82	1.26	105	15	111	27	76	14	10.7	4.4	11.2	4.6
Apical	8.27	1.47	105	19	171	34	75	11	15.6	5.3	17.3	7.5
Dorsal	9.41	1.33	139	21	143	25	101	18	8.7	2.5	10.6	4.8
/a:/	13.95	1.89	191	44	155	29	144	14	6.1	1.8	10.8	5.3
/i:/	11.13	1.54	224	38	101	26	157	39	5.1	1.7	13.8	7.9

TABLE III. Range of parameter values; Absolute values and percentage of mean difference for calculated to predefined parameter values by fitting synthetic data $diff_{syn}$ ($N=114$, synthetic data: corpus III); And absolute values and percentage of mean difference for procedure III to procedure II parameter values $diff_p$ ($N=100$, natural data: Corpus I).

Param.	Units	Range abs.	$Diff_{syn}$ abs.	$Diff_{syn}$ [%]	$Diff_p$ abs.	$Diff_p$ [%]
y_{tg}	[mm]	5.0–13.0	1.3	14.4	...	...
T_0	[ms]	100–300	8.2	5.3	12.6	8.4
t_{1on}	[ms]	...	7.6	...	12.7	...
t_{2on}	[ms]	40–100	3.4	4.8	5.1	7.2
t_{1off}	[ms]	60–150	9.5	9.1	4.2	4.0

relative difference around 15% indicating that target estimation by fitting is not very precise. Estimation of eigenperiod T_0, beginning and end of onset interval t_{1on} and t_{1off} (fitting procedure II) yields low relative differences around 5%. Beginning of offset interval t_{1off}, which indicates the end of the fit interval in the case of procedure II, is around 10%. (The accuracy of the parameter t_{2off} was not estimated, since in this case we have to synthesize complexes of three overlapping gestures.)

The parameters T_0, t_{1on}, t_{2on}, and t_{1off} can also be estimated by procedure III, taking into account the overlap with the preceding gesture. Fitting procedure III in addition estimates the end of the offset interval t_{2off} of the preceding gesture (not evaluated here). Since overlap is very important within a gestural theory, and since only the succession of fittings for all gestures leads to the full resynthesis of the movement traces of complete utterances, it is very important to evaluate the differences for eigenperiod T_0 and the time values t_{1on}, t_{2on}, and t_{1off} between procedure II and procedure III ($diff_p$). We estimated these differences by fitting natural data. In order to ensure that the full parameter range given in Table III is covered, we analyzed 20 tokens of each type of gesture of corpus I by applying procedures II and III. It can be seen from Table III that the parameter differences resulting from both procedures are less than twice the values of $diff_{syn}$, i.e., the measurement error for parameter estimation of the algorithm itself. That indicates that piecewise fitting as done by procedure II is sufficient for a rough evaluation of gestural parameters.

The fact that gestural overlap only slightly influences the estimated parameter values of the following gesture can be understood from the influence of the force fields on the articulator during an overlap of two gestures. Since the articulator is near the target of the preceding gesture, i.e., near the origin of the force field of this gesture (Figs. 2 and 3), the acceleration from the field of the preceding gesture is considerably lower than that from the force field of the following gesture during the interval of overlap. Therefore, the influence of the force field of the preceding gesture is relatively weak, if this force field dies out before the articulator moves far from the target of the preceding gesture.

2. Target estimation

In contrast to extreme displacement y_{ex}, gestural target position y_{tg} cannot be measured directly from movement data, since gestural targets lie beyond the extreme displacement points of any gestural trajectory. Target position can only be estimated indirectly from the shape of the articulatory displacement-time functions by taking into account, in particular, the main gestural velocity peak. Analysis of synthetic movement traces suggests that target estimation is not very accurate. As stated above, target estimation is made by only analyzing unreduced gestures, i.e., gestures within the most stressed syllable of an utterance. For these gestures the minimal articulator target distance, i.e., the distance between extreme displacement and target is low. Since we focused on articulatory movements in /i/-/a/ contexts, target positions are estimated only within the dimension of main movement component.

TABLE IV. Standard deviation (s.d.) for measured displacement maxima y_{ex}, standard deviation for calculated (fitted) target positions y_{tg}, and absolute mean values (ab) and standard deviation for the target to extreme distance y_{tg}-y_{ex} (corpus I, all speakers, all sessions; $N=1080$).

Gesture	y_{ex} [mm] s.d.	y_{tg} [mm] s.d.	y_{tg}-y_{ex} [mm] ab	y_{tg}-y_{ex} [mm] s.d.
Labial	1.10	1.33	1.09	0.97
Apical	0.65	1.60	2.23	1.78
Dorsal	1.23	2.33	2.63	2.31
/a:/	1.76	3.40	4.64	3.04
/i:/	1.02	1.35	0.53	0.96

The results for the accuracy of target estimation are given in Table IV. This table compares the kinematic variable extreme displacement y_{ex} with the estimated model parameter target position y_{tg}. It can be seen that standard deviation for target position is one to three times greater than the standard deviation for extreme displacement. This reflects the fact that the extreme displacement is directly measurable, while target position is not. It is remarkable that the standard deviation for target position is of the same order of magnitude as the difference of target position to extreme displacement y_{tg}-y_{ex}.

The high standard deviation of the extreme displacement for /a/-gestures can be explained by the fact that our tongue body receiver coil does not reflect the area of main constriction in case of the /a/. For this vowel, the main constriction occurs in the pharyngeal part of the vocal tract (Wood, 1979). Perkell and Nelson (1982) found lowest variance for target position of /i/ and /a/ in the region of the main constriction, perpendicular to the vocal tract walls, i.e., in the palatal region and vertical direction for /i/ and in the pharyngeal region and horizontal direction for /a/. This component of precise articulation need not result from precise muscular actions. It can be explained by anatomically based saturation effects (Fujimura and Kakita, 1979). Saturation effects may also be the reason for abrupt changes in articulator motion, which we found for all types of gestures except lip gestures. These abrupt changes always constitute the end of the fit interval for procedures I and II. For example, in the case of an apical closing gesture, saturation originates from the fact

TABLE V. Eigenperiod values T_0, phase values of peak velocity pha_{vp}, and relative articulator-target-distances at 180 and 360 deg yr_{180} and yr_{360} for all speakers and all sessions (corpus I, $N=1080$). ab: absolute values; s.d.: standard deviation.

	T_0 [ms]		Pha_{vp} [deg]		Yr_{180} [%]		Yr_{360} [%]	
Gesture	ab	s.d.	ab	s.d.	ab	s.d.	ab	s.d.
Labial	127	26	101	13	25.2	2.0	2.35	0.26
Apical	114	19	104	15	25.9	2.0	2.49	0.25
Dorsal	169	35	106	17	26.2	2.1	2.40	0.28
/a:/	240	50	108	17	26.2	1.8	2.34	0.26
/i:/	261	72	98	24	24.8	2.1	2.16	0.27

that the tongue blade is pushed against and restrained by the hard palate ("clipping" in our gestural production model; Kröger, 1993a, b).

Since the contact of the tongue tip or blade with the alveolar ridge or the hard palate for apical and dorsal closing gestures is clearly visible in the movement traces, the standard deviation of extreme displacement y_{ex} for lingual closing gestures should be zero. The actual nonzero standard deviation (Table IV) first results from the limited measurement precision of the device (e.g., tilting receiver coils) and from movements of the helmet position relative to the scull during a session. Second, the nonzero standard deviation results from the varying constriction *location* of the gesture-performing articulator. Due to our data reduction into one dimension (i.e., into the direction of *degree* of constriction), changes of constriction location cannot be detected but may lead to small changes of vocal tract wall position in the main movement direction. Therefore, more complex two-dimensional procedures for target estimation should be introduced for further investigations.

3. Estimation of gestural onset location and gestural eigenperiod

Mean values for eigenperiod and phase value of peak velocity estimated from natural data using procedure II are given for each type of gesture in Table V (corpus I) and Table VI (corpus II). At a first view these data could be interpreted in the following way: Eigenperiod differs for different types of gestures while the phase value of peak velocity is roughly constant (mean value is 102 deg). Thus eigenperiod clearly reflects the type of gesture (e.g., consonantal versus vocalic). If the phase value of peak velocity is taken as a measure for gestural onset interval location, its constancy would indicate that the gestural onset interval is localized in a similar (not necessarily identical) way for different types of gestures. But a closer analysis of the data including statistical analyses indicates that the situation is more complex.

TABLE VI. Eigenperiod values T_0, phase values of peak velocity pha_{vp}, and absolute time and phase value for end of gesture dt_{end}, pha_{end} for all speakers and all sessions for two vocalic gestures (long) /a:/ gesture. and (short) /a/ gesture) (corpus II, $N=360$). ab: absolute values; s.d.: standard deviation.

	T_0 [ms]		Pha_{vp} [deg]		dt_{end} [ms]		Pha_{end} [deg]	
Gesture	ab	s.d.	ab	s.d.	ab	s.d.	ab	s.d.
/a:/	239	42	101	13	156	33	350	80
/a/	259	35	96	11	78	19	207	32

Analyses of variance were performed for the gestural eigenperiod values of corpus I in order to evaluate the effects of (type of) gesture ($df=4,1079$), speaker ($df=2,1079$) and session ($df=1,1079$). The main effects were significant for gesture ($F=575.81$, $p<0.0001$), speaker ($F=30.78$, $p<0.0001$), and session ($F=8.65$, $p<0.01$). Two- and three-way interaction effects reached significance except for speaker/session interaction ($p>0.05$). Thus also the factors of speaker and session reached significance, but the amount of variance accounted for by the factor gesture is 63% for our data while the total amount of variance accounted for by all three factors (including all interactions) is only 8% higher. This indicates that the factor gesture is the most important one, but, up to a certain degree, the eigenperiod also reflects variation referring to speaker and session.

Post hoc Scheffé comparisons (Hays, 1988, p. 415ff) for different types of gesture show that the high significance level ($p<0.0001$) of the main effect gesture can only be reached for the difference of vocalic versus consonantal gestures and for the difference of the dorsal consonantal versus other consonantal gestures (labial and apical). The remaining independent comparisons between the vocalic gestures (/a:/ vs /i:/) and between the nondorsal consonantal gestures (labial versus apical) yielded lower significance levels ($p<0.001$ and $p<0.05$). Thus, for eigenperiod, three classes of gestures can be differentiated: Vocalic gestures, dorsal consonantal and nondorsal consonantal gestures.

Analyses of variance performed for the phase value of peak velocity for Corpus I showed significant main effects for gesture ($F=12.14$, $p<0.0001$), for speaker ($F=9.75$, $p<0.0001$), and for session ($F=10.06$, $p<0.005$), while most interaction effects were not significant. This indicates that the gestural velocity peak occurs at (slightly but significantly) different phase values for different types of gestures which results from the fact that different types of gestures exhibit slightly different velocity profiles. Again, the factors speaker and session reached significance but the total amount of variance accounted for by all three factors (including all interactions) is only 11% for these data.

Table V also gives mean values for the relative articulator-target distance at 180 and 360 deg. It should be noted that these measures refer to the hypothetical case of an undisturbed and temporally unlimited gesture (see the solid lines in Fig. 4). At a first view these values are roughly constant for all types of gestures. Mean values are around 25.7% for 180 deg and around 2.35% for 360 deg. Thus phase values may deliver a measure for the degree to which a gesture has been performed. But analyses of variance yielded significant main effects and significant interaction effects. Thus small but significant interspeaker- and intersession-differences occur. The total amount of variance

(including all interactions) accounted for by the three factors is only 18% for the 180-deg value and only 23% for the 360-deg value for these data.

Despite the fact that the mean values for the phase value of peak velocity as well as for the relative articulator-target distances at 180 and 360 deg are relatively stable for all gestures, the great amount of variance which cannot be accounted for by the factors analyzed above indicates that the calculation of phase values for a gesture is not very precise. Our fitting procedure gives only a rough estimate for the location of the phase scale for a gesture.

Beside eigenperiod and the phase value of peak velocity Table VI gives two measures for the time of extreme displacement, which indicates the end of the gesture as defined above by the starting conditions. The first variable is the absolute length of the time interval starting at the beginning of the gestural onset phase, while the second is the appropriate phase value. For both measures significant differences ($p<0.0001$) occur for the two types of vocalic gestures, i.e., German long versus short vowel gestures. Together with the relatively similar eigenperiod values for both types of gestures, this reflects the fact that both types of gestures mainly differ in their duration. It is remarkable that the phase values estimated by our quantitative approach are comparable to the release phase values given by Browman and Goldstein (1990).

III. GENERAL DISCUSSION

A quantitative model (six-parameter model) for fitting articulatory gestures has been developed. On the one hand, this model is based on current linguistic approaches (Browman and Goldstein, 1989) and, on the other hand, on a physical dynamic model for the analysis and description of speech movements. According to our detailed dynamic model, high-fit accuracy is obtained not only for displacement but also for velocity and acceleration. The model introduces time intervals of finite extent for gestural onset and offset. Onset and offset are seen as continuous and nonabrupt processes. Abrupt changes of dynamic gestural parameters are avoided. A main point, which differentiates our fitting procedure from others (e.g., from Browman and Goldstein, 1985; Smith *et al.*, 1993), is that our algorithm does not use predefined and fixed time windows for gestures. The estimation of gestural location is the *result* of our fitting procedure itself. Estimation of gestural onset as well as offset is one of the main features of this gestural analysis. The model is able to fit a great variety of articulatory movement shapes without violating the assumption of monotonicity for gestural movement shapes; i.e., each gesture represents a monotonically ascending or monotonically descending movement.

Fitting synthetic data shows that the onset interval location and eigenperiod can be estimated precisely. Fitting natural data indicates that the estimation of onset location is stable. Eigenperiod values differ significantly for consonantal and vocalic gestures, which is in agreement with hypotheses stated by Browman and Goldstein (1990). For consonantal gestures eigenperiod values differ with respect to the gesture-executing end-articulator. Gestural targets cannot be measured directly from the movement data, since by definition targets are never completely reached by the gesture-executing articulator. Targets are located beyond extreme gestural displacements in all cases and can be estimated only from the kinematics and underlying dynamics of the associated gestural time function. Thus our target estimation procedure leads only to a rough estimation of the target position.

The fitting algorithm introduced in this approach is a piecewise fitting of articulator movements, that results from the strong coupling of our dynamic model with a linguistic concept, i.e., the concept of discrete gestures with finite extent in time. The piecewise fitting may be a shortcoming since it makes automated data processing difficult. Interactive procedures are needed in order to obtain appropriate starting values, which makes the fitting time-consuming. But it is shown that fitting whole movement traces is possible (fitting procedure III). From a comparison of eigenperiod and onset estimation performed by different procedures, i.e., isolated fitting of a gesture (procedure II) compared with simultaneous fitting of two gestures (procedure III), it can be concluded that parameter estimation is possible by an isolated fitting that neglects overlap of adjacent gestures. This facilitates parameter extraction from movement data, since otherwise complex procedures for simultaneous parameter estimation for all gestures of an utterance would be necessary.

Within this approach gestural onset and offset are not modeled as discrete events, i.e., distinct time instants, but as continuous processes, i.e., time intervals for gestural onset and offset. The location of these intervals, i.e., the estimation of the overall location and shape of gestural strength functions or activation pulses for each concrete gesture of an utterance, is a main result of our fitting procedure.

It should be noticed that our measurements neither support nor contradict concepts of relative or absolute timing. Phase values are simply used as a measure for gestural magnitude. Since, from our viewpoint, eigenperiod is also a gesture-inherent parameter — in addition to target position — it is possible to measure gestural phasing by phase values as well as by absolute time values. But this six-parameter model delivers a concrete method for calculating intrinsic gestural phase values. Our measurements indicate that the relative articulator-target-distance mainly depends on phase. Therefore, a phase value describes the degree to which a gesture has been executed, i.e., the instantaneous gestural magnitude. Furthermore, phase may be an important concept for describing the duration of long versus short vowel gestures (see analysis of Corpus II), intergestural timing in stress contrasts, and speech tempo variations. In further experiments it is necessary to evaluate these kinds of linguistic influences using this quantitative model.

A long-termed goal of this work is the implementation of this quantitative six-parameter model as the basis for the control module of an articulatory synthesizer. This can improve the naturalness of generated articulator movements in comparison to gestural movements based on step-rectangular-shaped strength functions. Since our extended dynamic model for speech gestures is capable of fitting articulatory movement traces with high accuracy, it can be assumed that the model is realistic and that the extracted dy-

namic parameters are significant and useful for describing speech movements.

ACKNOWLEDGMENTS

This work was supported by the German Research Council (DFG) Grant: Kr 1439/2-1 and He 434/22-1 and, in part, by ESPRIT Project No. BR-6975 (SPEECH MAPS).

Browman, C. P., and Goldstein, L. (**1985**). "Dynamic modeling of phonetic structure," in *Phonetic Linguistics. Essays in Honor of Peter Ladefoged*, edited by V. A. Fromkin (Academic, New York), pp. 35–53.

Browman, C. P., and Goldstein, L. (**1989**). "Articulatory gestures as phonological units," Phonology **6**, 201–251.

Browman, C. P., and Goldstein, L. (**1990**). "Tiers in articulatory phonology, with some implications for casual speech," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. E. Beckman (Cambridge U.P., Cambridge), pp. 341–376.

Flanagan, J. R., Ostry, D. J., and Feldman, A. G. (**1990**). "Control of human jaw and multi-joint arm movements," in *Cerebral Control of Speech and Limb Movements*, edited by G. R. Hammond (North-Holland, Amsterdam), pp. 29–58.

Fujimura, O., and Kakita, Y. (**1979**). "Remarks on quantitative description of the lingual articulation," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhmann (Academic, London), pp. 17–24.

Hays, W. L. (**1988**). *Statistics* (Holt, Rinehart, and Winston, Fort Worth, TX), 4th ed.

Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (**1986**). "The dynamical perspective on speech production: data and theory," J. Phon. **14**, 29–59.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. (**1985**). "A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling," J. Acoust. Soc. Am. **77**, 266–280.

Kröger, B. J. (**1993a**). "A gestural production model and its application to reduction in German," Phonetica **50**, 213–233.

Kröger, B. J. (**1993b**). "A gestural approach for controlling an articulatory speech synthesizer," 3rd Eur. Conf. Speech Commun. Technol. Proc. (Berlin) **3**, 1903–1906.

Perkell, J. S., and Nelson, W. L. (**1982**). "Articulatory targets and speech motor control: a study of vowel production," in *Speech Motor Control*, edited by S. Grillner, B. Lindblom, J. Lubker, and A. Persson (Pergamon, Oxford), pp. 187–204.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Mattheis, M. L., Garabieta, I., and Jackson, M. T. T. (**1992**). "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096.

Perrier, P., Laboissière, R., and Eck, L. (**1991**). "Modelling of speech motor control and articulatory trajectories," Proc. XIIth Int. Congr. Phonet. Sci. **2**, 62–65.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1992**). *Numerical Recipes in C* (Cambridge U.P., Cambridge), 2nd ed.

Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecol. Psych. **1**, 333–382.

Schönle, P., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., Conrad, B. (**1987**): "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," Brain Lang. **31**, 26–35.

Shigenaga, M., and Ariizumi, H. (**1977**). "Articulatory movements by rule," in *Articulatory Modeling and Phonetics*, edited by R. Carre, R. Descout, and M. Wajskop (G.A.L.F., Group de la Communication parlée, Grenoble), pp. 193–202.

Shirai, K. (**1993**). "Estimation and generation of articulatory motion using neuronal networks," Speech Commun. **13**, 45–51.

Smith, C. L., Browman, C. P., McGowan, R. S., and Kay, B. (**1993**). "Extracting dynamic parameters from speech movement data," J. Acoust. Soc. Am. **93**, 1580–1588.

Sonoda, Y. (**1977**). "Estimation of dynamic characteristics of articulatory movements," in *Articulatory Modeling and Phonetics*, edited by R. Carre, R. Descout, and M. Wajskop (G.A.L.F., Group de la Communication parlée, Grenoble), pp. 213–222.

Tuller, B., Shao, S., and Kelso, J. A. S. (**1990**). "An evaluation of an alternating magnetic field device for monitoring tongue movements," J. Acoust. Soc. Am. **88**, 674–679.

Vatikiotis-Bateson, E., Hirajama, M., and Kawato, M. (**1991**). "Neural network modelling of speech motor control using physiological data," Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (PERILUS) **14**, 63–68.

Wood, S. (**1979**). "A radiographic analysis of constriction location for vowels," J. Phon. **7**, 25–43.