Original Paper

Phonetica 1993; 50: 213–233

*Bernd J. Kröger*

Institut für Phonetik,
Universität Köln, BRD

# A Gestural Production Model and Its Application to Reduction in German

**Abstract**
A quantitative speech production model has been computer-implemented based on specifications of gestures as its input. Articulatory gestures serve the central role of phonological/phonetic units in speech organization. The present model is not based on task dynamics, but it still assumes a critically damped linear second-order system. The capability of this model is partially demonstrated by its application to reduction phenomena in conversational speech in German. Explanations in terms of timing and magnitude alterations of gestures in different articulators use only a few general principles for a variety of apparent segmental alterations. Some salient examples have been synthesized based on the current model, which also contributed to the evaluation of the parameter values involved in sample gestures.

## Introduction

A gestural production model including an articulatory-acoustic component has been developed which addresses the theoretical problem of converting phonological units into articulatory movements and which is practically capable of producing synthetic speech. The first step was the implementation of an aerodynamic-acoustic model for the vocal folds [Kröger, 1990a] and for the vocal tract [Kröger, 1990b]. The second step was adapting the IPKöln articulatory model [Heike, 1979, 1980, 1989]. As the third step, spatial targets for vowels and consonants were estimated and first rules for the timing between different articulators were established: A segmental rule component comprising only very

Received:
December 23, 1991
Accepted:
April 19, 1993

Bernd-J. Kröger
Institut für Phonetik
Universität Köln, Greinstrasse 2
D-50939 Köln (FRG)

primitive rules was developed [Kröger, 1992]. It could be shown that this component is able to control an articulatory synthesizer with a very small inventory of underlying articulatory targets. This inventory corresponds to a sound inventory which is only slightly greater than the phoneme inventory of German, since a concept of articulatory underspecification leads to the needed allophonic variation.

But this model shows the typical shortcomings of segmental concepts: (1) Explicit rules for segment durations must be implemented. (2) A suprasegmental rule component is like an annex to the segmental rules and not a naturally integrated part of the model. (3) If the speech rate is increased, segmental changes related to reduction must be explicitly incorporated as a set of phonological rules leading to an altered segment chain. These shortcomings are strongly related to the segmental-phonological background of this model and can be overcome only by changing its underlying principles. Thus as the fourth step a nonsegmental production model using the concept of gestures as described by Browman and Goldstein [1987] was developed.

The theory of articulatory phonology developed by Browman and Goldstein [1986, 1987, 1988, 1989, 1992] assumes that relatively stable units of articulatory movements called *gestures* serve as the basic elements of speech production. For example in order to produce the syllable /ba/, two gestures occur: a labial full-closing gesture and a dorsal-pharyngeal constriction-forming gesture. Additionally, information about the timing or *phasing* between these gestures is needed. The consonantal gesture must be activated earlier than the vocalic gesture and the vocalic gesture must be activated during the time interval of lip closure, i.e. before the activation of the labial full-closing gesture ends. This *gestural overlap* in the time domain is a very important feature of articulatory phonology.

The recognition of articulatory gestures as stable and basic units of speech production and of overlap among gestures is not new in phonetic theory. For example Joos [1948, pp. 109–126] postulates 'innervation waves', each causing definite muscular actions and definite articulator movements. He introduced an 'overlapping innervation wave theory' to explain coarticulatory influences between adjacent sounds. Stetson [1951, p. 43] states that syllable-initial consonants ('releasing consonants') overlap the vowel process and that these consonants do not add to the duration of the syllable. The main contribution of Browman and Goldstein [1987] is to introduce the gesture as the basic unit of articulatory movements associated with a concrete and quantitatively defined concept of intergestural timing. The dynamics of gestures is described by a critically damped mass-spring equation leading to the parameters mass, stiffness, and equilibrium or rest position which are reset for each gesture. In addition, phasing rules [Browman and Goldstein, 1987, pp. 12–14] determine the intergestural timing.

But this quantitative description is not without problems. Firstly, the assumption of stable phase relations between gestures and the assumption that the beginning of the activation of a gesture is triggered according to a certain phase value of another gesture is not accepted generally [Lubker, 1986; Fujimura, 1990]. But the phase description for gestures [Browman and Goldstein, 1987] is useful if phase is taken as a measure for the degree to which a gesture has been executed. Secondly, the term stiffness might suggest that the mass-spring system concretely models the dynamics of each articulator. To avoid this problem [Fujimura, 1986b, 1990] stiffness in combination with the assumption of unit mass is replaced by eigenperiod in this article. So we use eigenperiod and phase to effectively describe gestures and their timing quantitatively.

Reduction is a very common phenomenon in speech. Nearly every phonetic realization of words or phrases in connected speech is reduced to a certain extent. A possible classifying parameter for the various reduced forms of a phrase is the degree of phonetic alteration in comparison with a slowly and precisely uttered form, depending on the position in the sentence, the importance of the information carried, and the communication situation in which the utterance is produced. Obviously, the degree of reduction depends on speaking rate: The higher the speaking rate (i.e. the shorter the time interval in which the utterance has to be produced) the higher the degree of reduction. The gestural framework offers two different models for reaching high speaking rates: (1) Decrease in gestural eigenperiod for all gestures but constant intergestural phasing. (2) Changes with respect to intergestural constellation: in particular increase in gestural overlap and decrease in gestural extension retaining constant gestural eigenperiod.

Model 2 leads to several nonlinear transformation processes due to changes in speech rate, as discussed by Gay [1981]. It leads to a variety of segmental changes [for German see Kohler, 1990]. Model 1 predicts higher articulator velocities. All segment durations should decrease proportionally and no discrete segmental change should occur. A very common phenomenon which contradicts this model is undershoot [Lindblom, 1964, 1983]: there is not sufficient time to reach articulatory target configurations in fast speech since the velocity of the movements towards them is limited. Lindblom [1983, pp. 226–231] developed a gestural theory which is able to explain articulatory undershoot as a result of limits in gestural force and gestural duration. Constant gestural force, which corresponds to constant gestural eigenperiod in terms of our concept, leads to stable movement patterns, i.e. to invariant transient portions of the gesture-induced articulator movements. Fujimura [1981, 1986a] called such invariant transient movement patterns 'icebergs' or 'elementary articulatory gestures' and he stated that primarily the *inter*gestural timing changes according to context, stress, and speaking rate while the *intra*gestural movement patterns remain stable. Changes in the timing or phasing of gestures as a consequence of continuously increasing speaking rates were also measured by Kelso et al. [1986, pp. 50–52]. Later, phase changes were attributed to linguistic influences [Nittrouer et al., 1988] and especially to stress [Edwards et al., 1991]. According to these experimental results, the second model above seems appropriate for modelling reduction.

Kohler [1990] presented a set of generative rules which describe the segmental changes leading to various degrees of reduction in German and he states that a series of segmental changes which occur for a given phrase are the consequence of few underlying 'processes' which can be divided into different 'groups according to speech production criteria' [Kohler, 1991b, p. 187]. It will be shown in this study that our gestural production model is capable to specify these gestural alteration processes. For his example *mit dem* in German, which exhibits a number of reduced segmental variants, many of them can be predicted by two types of gestural alteration processes proposed by Browman and Goldstein [1987, 1989]: (1) the increase in gestural overlap and (2) the decrease in gestural extension. A third type of gestural alteration process, the gesture-executing articulator swap, is proposed in this article to explain all reduced forms of this German expression cited by Kohler.

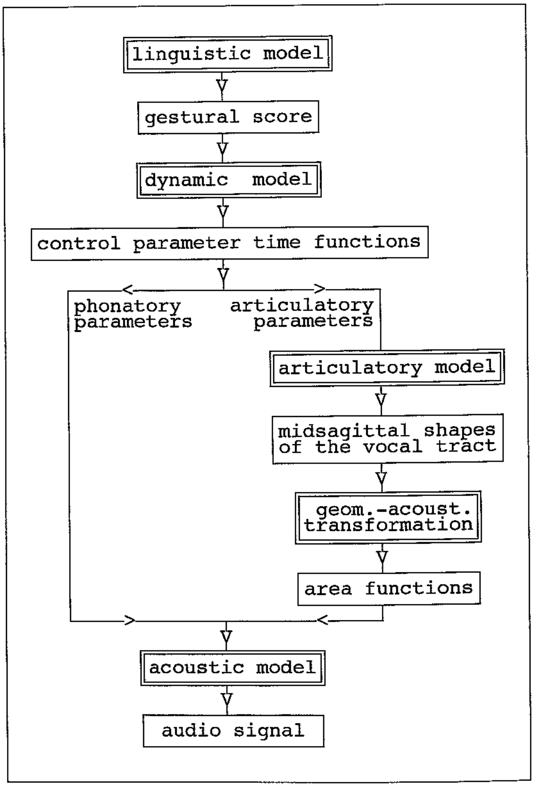

**Fig. 1.** Layers (single-lined rectangles) and modules (double-lined rectangles) of the production model.

## The Gestural Production Model: An Overview

The organization of the production model is given in figure 1. The linguistic model generates the gestural score by specifying the gestures and all gestural descriptors. The dynamic model produces continuous control parameter time functions. The phonatory control parameters (glottal aperture, pulmonary pressure, cord tension) can be applied directly to the acoustic model. The articulatory control parameters (controlling the lips, tongue, and velum) must be transformed by the articulatory model into vocal tract shapes and then by the geometric-acoustic transformation into acoustically relevant area functions. The acoustic model generates the audio signal.

## From Articulator Movements to the Acoustic Signal: The Articulatory-Acoustic Part of the Production Model

### *The Articulatory Model*

An articulatory model defines parameters describing possible vocal tract shapes. Different kinds of parameterizations can be useful depending on the task for which the articulatory model is used. On the one hand, the parameterization can be directly related to the acoustically relevant vocal tract shape itself. In this case possible parameters are length of the tube and location and degree of the main tract constrictions [Stevens and House, 1955; Fant, 1960, pp. 71–90]. On the other hand, the parameterization can be related to articulators, each forming its own dynamic subsystem, and

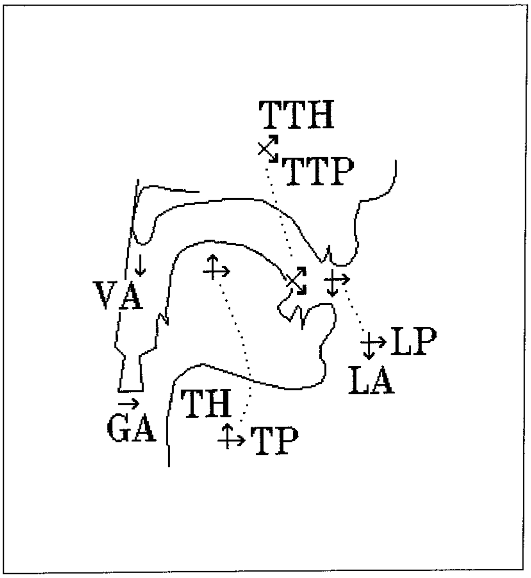

**Fig. 2.** Midsagittal view of the vocal tract shape produced by the articulatory model. The arrows indicate the regions mainly influenced by the control parameters and the directions of major change. The control parameter symbols are listed in table 1.

relative coordinates can be used describing the location of an articulator with respect to the location of other articulators (e. g. lip or tongue body coordinates relative to the jaw position) [Mermelstein, 1973]. But both concepts have shortcomings. While tract shape-related concepts cannot model the physical system responsible for the articulator dynamics, the articulator-related concepts can lead to an intricate and too indirect description of the tract shape. For example in the latter case the constriction degree formed by the tongue tip is a function of tongue tip, tongue body, and jaw position. In order to avoid these shortcomings we use four uncoupled and independent articulators which are controlled by parameters directly describing the degree and location of vocal tract constrictions (fig. 2). Three articulators, the lips, the tongue tip, and the tongue body mainly control the vocal tract shape [Coker, 1967] while the fourth supraglottal articulator, the velum, mainly controls the degree of nasal tract coupling.

The midsagittal view of every vocal tract shape generated by our model is defined numerically by the x/y coordinates of 128 sample points each. Subsets of these 128 sample points define the midsagittal projection of the tract cavity surface of each articulator. This projection in the region of an articulator for a given control parameter value is calculated by linear interpolation from defined extreme configurations for that articulator. This linear interpolation must be done for each coordinate of each sample point within the subset of the articulator concerned. If the articulator is controlled by one parameter, two extreme configurations are defined, one corresponding to the minimal and one to the maximal value of the control parameter (e. g. VA=–100 and VA=+100 in the case of velum; table 1). If the articulator is controlled by two control parameters, four extreme configurations are necessary, each corresponding to an edge point of the two-dimensional parameter space (e. g. {TH=–100, TP=–100}, {TH=–100, TP=100}, {TH=100, TP=–100}, and {TH=100, TP=100} in the case of tongue body).

### *The Control Parameters*

The production model operates on the basis of seven articulatory control parameters defined by the articulatory model and on the ba-

**Table 1.** List of control parameters used in the model, the range of control parameter values, and their equivalent articulator positions

| Symbol | Name | Range of values | Equivalent articulator position |
|---|---|---|---|
| VA | velic<br>aperture | –100<br>0<br>100 | strongly raised (closure)<br>raised (closure)<br>lowered (opening) |
| LP | lip<br>protrusion | 0<br>100 | spread (unrounded)<br>protruded (rounded) |
| LA | lip<br>aperture | 0<br>100 | closed<br>opened |
| TH | tongue<br>height | –100<br>100 | lowered (pharyngeal)<br>raised |
| TP | tongue<br>position | –100<br>100 | back (velar)<br>fronted (palatal) |
| TTH | tongue tip<br>height | 0<br>100 | no elevation<br>occlusion |
| TTP | tongue tip<br>position | –200<br>–50<br>0<br>100 | retroflex<br>postalveolar<br>alveolar<br>dental |
| GA | glottal<br>aperture | –400<br>0<br>600 | strongly closed (glottal stop)<br>closed (normal phonation)<br>widely opened |
| CT | cord<br>tension | 0<br>200 | no tension<br>high tension |
| PR | pulmonary<br>pressure | 0<br>200 | no pressure<br>high pressure |

The extreme values for the control parameters are chosen arbitrarily. The range of each scale reflects the required degree of fineness, since the articulatory model handles integers.

sis of three phonatory control parameters defined by the glottis model of the acoustic model. All control parameters are listed in table 1. The three articulators – lips, tongue tip, and tongue body – are each controlled by two parameters, one describing the degree of constriction (or aperture in the case of the lips), the other describing the location of constriction (or protrusion in the case of the lips) whereas the velum is controlled by one parameter, the velic aperture (fig. 2, table 1). In contrast to glottal aperture, the two other phonatory control parameters – pulmonary pressure and vocal fold tension – must be varied only for modelling suprasegmentals [Kröger et al., 1991]. In this study constant values were chosen for these two parameters in such a way that a state of normal phonation can be produced.

*The Acoustic Model*

The acoustic model comprises a vocal tract model and a self-oscillating glottis model [Kröger, 1990a, b]. The vocal tract model is based on the Kelly-Lochbaum [1962] line model, extended by acoustically and aerodynamically important loss phenomena as described by Liljencrants [1985]. The pressure and the volume flow of the air in the glottis-lip direction in the vocal tract are calculated for each time instant. Turbulent noise is automatically generated and inserted into the line, if the pertinent conditions – narrow constriction and sufficiently high volume flow within this constriction – are fulfilled. The self-oscillating glottis model is based on the Ishizaka-Flanagan [1972] model. It simulates the dynamic behavior of the vocal cords (e.g. cord vibration during phonation) and calculates the instantaneous glottal area. In contrast to this area the control parameter glottal aperture (table 1) defines the slowly varying glottal rest area, i.e. the mean distance of the vocal folds according to abduction and adduction, and it controls for example the presence or absence of voicing.

## The Generation of Articulator Movements: Gestures and Their Coordination

On the one hand, gestures can be seen as *basic phonological units*. They define discrete categories like 'labial closure' or 'velic opening'. On the other hand, gestures are *units of action* [Browman and Goldstein, 1992]. They determine the continuous movements of articulators. In our quantitative production model the articulator movements are defined if the set of gestural descriptor values is specified for all gestures corresponding to a given utterance. Table 2 lists selected gestures for German. We differentiate three types of gestures: vocalic gestures (e.g. eedo, oodo, isdo, usdo), consonantal gestures (e.g. fcla, fcap, fcdo, ncal, ncpo), and opening gestures (opgl, opve).

Figure 3 shows selected control parameter time functions and the oscillogram of the synthetic speech signal for /mʊs/ ('must' in German). The control parameter time functions illustrate the movements of the articulators. Each gesture is activated only in a distinct time interval, i.e. its activation interval, represented by the horizontal extent of the shaded boxes in figure 3. Table 3 gives a full specification of the gestural score and figure 4 illustrates the gestural phasing. If a gesture is activated, the corresponding articulator executes a goal-directed movement towards a spatial target which is represented by defined control parameter values. If no gesture is active for an articulator a movement towards its inherent neutral position occurs. (The production state corresponding to the neutral positions of all articulators defines a schwa-like vocal tract configuration and normal phonation.)

One important feature of the gestural approach is that vowel gestures are in an immediate succession without a gap in fluent speech (see e.g. fig. 7b). They act as a 'ground' to consonantal 'figures' [Browman, 1991]. The articulatory movements resulting from these series of vocalic gestures are comparable to the 'vocalic base function' in Fujimura's [1992] C/D model or to the 'vowel component' in Öhman's [1967] model. Generally, in fluent speech, consonantal gestures are phased with respect to those time instants in which a vocalic gesture starts or ends, and thus they are completely overlapped by vocalic gestures.

*The Quantitative Dynamic Model of the Gesture*

The model of a critically damped linear second-order system is used for the quantitative description of the dynamics of a gesture

**Table 2.** List of selected gestures for German, their symbols, the control parameters involved, and values for targets and clipping

| Symbol | Name | Control parameter | Target value | Clipping value |
|---|---|---|---|---|
| fcla | labial full-closing | LA | –20 | 0 |
| fcap | apical full-closing | TTH | 120 | 100 |
| fcdo | dorsal full-closing | TH | 120 | 100 |
| ncal | alveolar near-closing | TTH | 120 | 96 |
| | | TTP | 0 | – |
| ncpo | postalveolar near-closing | TTH | 120 | 96 |
| | | TTP | –50 | – |
| opgl | glottal opening | GA | 400 | – |
| opve | velic opening | VA | 100 | – |
| eedo | dorsal /e:/ | TH | 60 | – |
| | | TP | 90 | – |
| oodo | dorsal /o:/ | TH | 40 | – |
| | | TP | –100 | – |
| isdo | dorsal /ɪ/ | TH | 70 | – |
| | | TP | 70 | – |
| usdo | dorsal /ʊ/ | TH | 60 | – |
| | | TP | –60 | – |

The dashes indicate the absence of clipping.

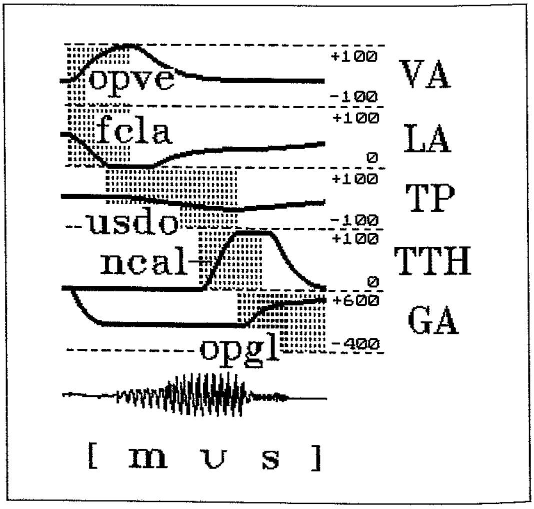

**Fig. 3.** Control parameter time functions (thick lines), gestural activation intervals (shaded areas), and the synthetic speech signal for /mʊs/. (The glottal opening gesture is overlapped by a postphonatory opening gesture.)

**Table 3.** Detailed specification of the gestural score of /mʊs/

| Gesture: | fcla | opve | usdo | | ncal | | opgl |
|---|---|---|---|---|---|---|---|
| Control parameter: | LA | VA | TH | TP | TTH | TTP | GA |
| Target value, – | –20 | 100 | 60 | –60 | 120 | 0 | 400 |
| Clipping value, – | 0 | – | – | – | 96 | – | – |
| Eigen-period, ms | 80 | 80 | 250 | | 80 | | 60 |
| Release phase, degrees | 290 | 290 | 200 | | 290 | | 290 |
| Association phase, degrees | 180 | 0 | – | | 180 | | 0 |

No further information is needed in order to produce this speech sample. The dashes for clipping values indicate absence of clipping; the dash for the association phase indicates that this vocalic gesture is not subordinate. A graphic display of this gestural score, including association lines, is given in figure 4.

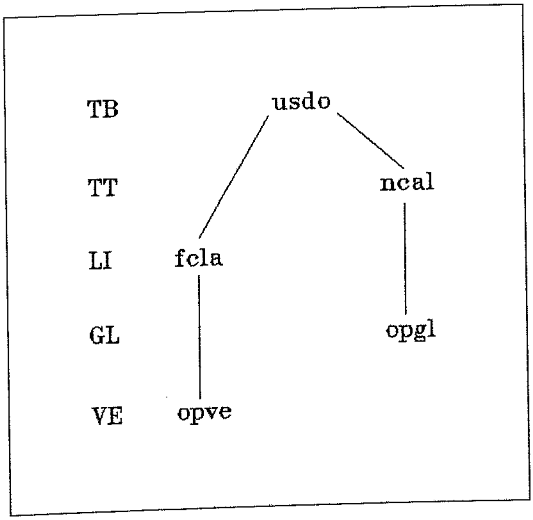

**Fig. 4.** Display of the gestural score of /mʊs/: Gesture symbols and association lines. The gestures are ordered with respect to articulatory tiers for tongue body (TB), tongue tip (TT), lips (LI), glottis (GL) and velum (VE). The lines indicate which gestures are associated with each other.

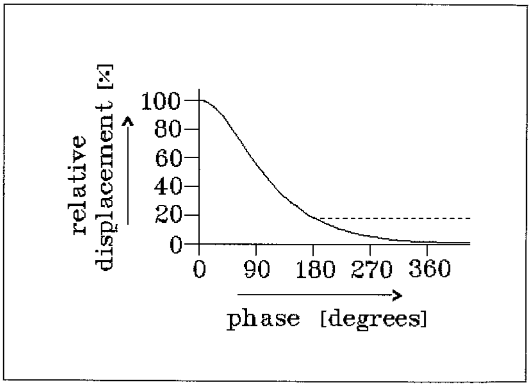

**Fig. 5.** Time function of an articulator movement if a gesture is activated for this articulator and if the initial articulator velocity is zero. The dashed line indicates clipping. Abscissa: phase (time relative to the eigenperiod of the gesture in degrees). Ordinate: displacement of the articulator from the gestural target relative to its initial displacement in percent.

[Saltzman, 1985; Browman and Goldstein, 1987]. This description can be directly applied to the gesture-executing articulator and its control parameter(s) since the former is independent of other articulators and since the latter are defined closely related to the tract shape. But it must be emphasized that the gesture – not the articulator – is described by the critically damped second-order system. Figure 5 shows the time function of an articulator movement if a gesture is activated for this articulator and if the initial articulator velocity is zero (for the mathematical description see 'Appendix'). The movement pattern for the relative displacement of the articulator from its gestural target is an asymptotic time function, never reaching zero displacement: the articulator never reaches the intended gestural target totally. In figure 5 relative scales are used for both axes. Displacement values are relative to the initial displacement. Time values are relative to the eigenperiod, i.e. to the duration of a single cycle of the oscillation of this linear second-order system if damping is removed. This leads to a definition of a phase scale for each gesture with different enlargement factor for gestures with different eigenperiod values. So a gesture with a lower eigenperiod value reaches a defined (small) articulator-target distance faster than a gesture with a higher eigenperiod value.

The model described so far produces articulatory movements in which the articulator-target distances decrease monotonically. But in order to produce plosives or fricatives, the supraglottal constriction gesture must exhibit a temporal interval with a constant degree of constriction. A constant degree of constriction can be produced if the gestural time function of an articulator movement is clipped, i.e. if the range of the gesture-induced articulator movement is limited, whereas the gestural target lies beyond this range limit. This 'clipping' is displayed in figure 5 by the dashed line; in figure 3 it occurs for the labial and apical closing gestures. The physiological origin of clipping is the contact of the articulator with its counterpart (e.g. lips or vocal folds) or the contact of the articulator with the vocal tract walls (e.g. tongue body with the palate) during the activation of these gestures. Figure 6 shows articulographic data of an apical full-closing gesture. The vertical tongue tip displacement is abruptly fixed by the collision of the tongue tip with the alveolar ridge whereas the vertical displacements of the tongue blade and the jaw do not show this behavior.

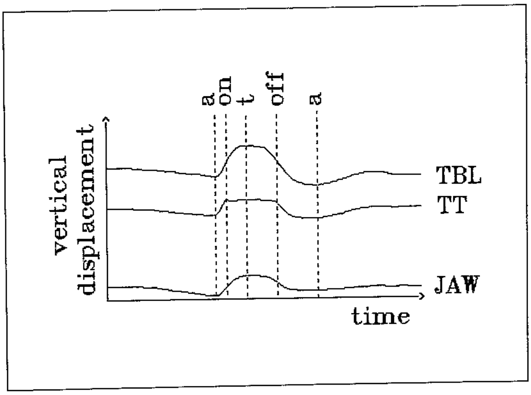

**Fig. 6.** Articulatory data of /ata/ obtained by the Articulograph AG-100 (Carstens Medizinelektronik GmbH, Göttingen, FRG). The diagram shows vertical displacement of jaw (JAW), tongue tip (TT), and tongue blade (TBL) as functions of time. The vertical dashed lines indicate the maximal articulatory displacements (returning points) for both /a/ sounds and /t/ as well as onset (on) and offset (off) of the tongue tip contact.

The clipping values, which describe the range limit quantitatively, and the target values are arranged in such a way that the portion of the gesture exhibiting the constant constriction starts at about 180 degrees (fig. 5). For all gestures including those without clipping we introduce the following rule of thumb: A rapid articulator movement towards the gestural target, i.e. the transient portion of a gestural movement, takes place at phase values below 180 degrees whereas the quasi-steady-state portion in which the articulator is near the gestural target (relative articulator-target distance is lower than ≈20%) appears at phase values above 180 degrees ('Appendix', table A1). The longer the gesture is activated above 180 degrees, the more quasi-steady-state portion of the gesture is produced. No consonantal closure is produced if the gestural activation ends below 180 degrees for the accompanying gesture.

*Gestural Descriptors*

Gestural descriptors define (1) the control parameter(s) on which the gesture operates, (2) the target, (3) the clipping value, (4) the eigenperiod, (5) the release phase, and (6) the association phase for each gesture. For instance the dorsal /e:/ gesture (table 2) specifies the time functions of tongue height and tongue position. The target is 60 for tongue height and 90 for tongue position (in the model arbitrary values are used; table 1). If a gesture is clipped, the range limit of the pertinent control parameter time function is defined by a clipping value. For instance, the labial full-closing gesture operates on the control parameter lip aperture. The target is –20, but the gesture is clipped at 0. Values for the gestural descriptors eigenperiod, release phase, and association phase are given in table 3. The eigenperiod determines roughly the articulator velocity and, together with the release phase, the duration of a gesture (the length of its activation interval). The association phase determines the intergestural coordination.

*The Coordination among Gestures: Gestural Phasing*

The association phase values of all gestures fix the intergestural constellation of intergestural phasing. Additionally, association conventions are needed, defining which gesture has to be phased with respect to which other gesture. These general conventions are [Browman and Goldstein, 1987, pp. 11–16] (fig. 4): (1) each vocalic gesture is phased with respect to the end of the preceding vocalic gesture; (2)

the first consonantal gesture of a consonant cluster is phased with respect to the beginning of the syllable-defining vocalic gesture if the cluster is syllable-initial, and with respect to its end if the cluster is syllable-final; (3) non-first consonantal gestures of a consonant cluster are phased with respect to the end of the preceding consonantal gesture within this cluster, and (4) opening gestures are phased with respect to the end of the pertinent consonantal gesture. These conventions define the association line as a time instant (e.g. 'end of the pertinent consonantal gesture'). This time instant has to coincide with that defined by the association phase value, which refers to the phase scale of the associated gesture.

### *The Dynamic Model*

The set of continuous control parameter time functions calculated by the dynamic model (fig. 1) describes the movements of all articulators. In our simplified model, differing from the task dynamic model [Saltzman, 1985], time functions for control parameters are independently computed without considering interactions among them. Within each gestural activation interval, control parameter time functions can be calculated from the gestural descriptor values together with the initial articulator-target distance and the initial articulator velocity for each gesture.

Two effects stemming from interarticulatory dependence are taken into account in our model: (1) According to the vowel-induced jaw height the neutral position for the lips changes if (dorsal) vocalic gestures are active. (2) According to the tongue tip-tongue body dependency, the neutral position of the tongue tip is a function of the instantaneous tongue body position. While (1) is modelled by varying the lip control parameter values for the neutral position, (2) is modelled within the articulatory model itself: the neutral tongue tip contour (constantly defined by {TTH=0, TTP=0}) is calculated for each time instant using the instantaneous tongue body contour.

### *Parameter Estimation*

The values for the gestural targets and for the clipping are strongly dependent on the articulatory model used. Vocalic target values were estimated perceptually using our articulatory synthesizer [Kröger, 1990a, b]. Consonantal target values for the place of the constriction were defined simply as the center of the pertinent region defined by the articulatory model (e.g. center of hard palate, soft palate). The clipping value, which defines the degree of constriction, equals an extremum within the range of the pertinent control parameter in the case of plosives (LA=0, TTH=100, and TH=100; table 1). In the case of fricatives this value is estimated by the acoustical criterion of maximal noise amplitude. Consonantal target values for the degree of constriction always lie beyond the threshold value of the pertinent control parameter. These target values are determined by the condition that the clipping values has to be reached roughly at 180 degrees.

Despite the fact that a quantitative gestural production model [Browman and Goldstein, 1987] has received much attention, there is no comprehensive attempt to estimate dynamic parameters from articulatory measurement data according to this particular model, i.e. there is no attempt to estimate eigenperiod, release phase and association phase. Some studies [for references see Smith et al., 1991] use the linear second-order system as a model for the dynamics of speech articulators but not explicitly the *critically damped* linear second-order system. Assuming release phase values around 300 degrees [Browman and Goldstein, 1987, pp. 11–16], eigenperiod values were estimated for our model by comparing segment durations of the synthetic speech signal to those naturally produced. Like Browman and Goldstein [1987] we used different release phase values and different eigenperiod values for vocalic and for consonantal gestures.

The intergestural phasing rules used in our model are approximately the same as those given by Browman and Goldstein [1987]. Only one major difference occurs. The association phase value for consonantal constriction gestures in our model is 180 degrees while Browman and Goldstein [1987] use a higher value. First listening tests with CV stimuli using the Browman and Goldstein value yielded a perceptual effect of heavy diphthongization for intended monophthongal vowels.

## Reduction Processes for German in the Gestural Model

In the gestural framework a wide variety of differences between the canonical pronunciation of words and their realization in fluent speech is supposed to be the result of two simple types of gestural alteration processes: (1) increase in overlap among gestures, and (2) reduction in magnitude (i.e. decrease in extension) of gestures in time and space [Browman and Goldstein, 1987, p. 17, 1989, p. 214]. Increase in overlap among two gestures is obtained by increasing the association phase value of the second or following gesture. Decrease in extension of a gesture is obtained by decreasing its release phase value.

Decreasing the release phase of a gesture leads to a decrease in gestural activation interval length. Consequently the minimal relative articulator target distance increases, i.e. the amplitude of the articulator movement decreases. Thus, a decrease of release phase leads to a temporal *and* to a spatial decrease of the gestural extension. This decrease in gestural extension can have qualitative effects, e.g. a centralization of vowel quality from [e:] to [ə].

Both processes, i.e. increase in gestural overlap or decrease in gestural extension, lead to the same result, i.e. to the same control parameter time functions, if the gestures involved are equal (e.g. two apical full-closing gestures), and if in addition the quasi-steady-state portions of both gestures overlap (in the case of consonantal gestures: the closure intervals produced by both gestures overlap). In this case both gestures become one single large gesture and increasing the association phase value of the second gesture leads to the same result as decreasing the release phase value of the preceding or first gesture. Both processes result in a decrease in extension of this large blended gesture. (If two overlapping gestures act on the same articulator as is for example the case for gestures of equal type, overlapping can also be called 'blending' [Browman and Goldstein, 1987, p. 18].)

### *The Example mit dem Boot: On Achieving the Fully Reduced Form*

In the German phrase *mit dem Boot* ('with the boat', 'by boat') the sequence *mit dem* can be affected by several discrete segmental changes. A list of reduced forms of *mit dem* is given by Kohler [1990, 1991a, b]. Trying to apply the two types of gestural alteration processes stated above, we found that two concrete gestural alteration processes are sufficient to produce the fully reduced form [mɪmbo:tʰ] [Kohler, 1990, p. 83]: (1) decreasing the extension of the dorsal /e:/ gesture, and (2) decreasing the extension of the glottal opening gesture of /mɪt/ (fig. 7). Firstly, the gaps between the activation intervals of the vowel gestures are eliminated (fig. 7b). A blending of the apical full-closing gesture of /mɪt/ and /de:m/ occurs as a byproduct. The glottal opening gesture of /mɪt/ is totally hidden by this blended apical full-closing gesture and consequently the acoustic noise burst of /t/ disappears. Secondly, the extension of the dorsal /e:/ gesture is reduced (decrease of the release phase value of this gesture to 200 degrees), which results in a change of the vowel [e:] to [ə] (fig. 7b). Further decrease in extension of the dorsal /e:/ gesture (decrease of release phase to 100 degrees) leads to an elision of this vowel. In this reduced form (fig. 7c) the dorsal /e:/ gesture still exists, but it is totally

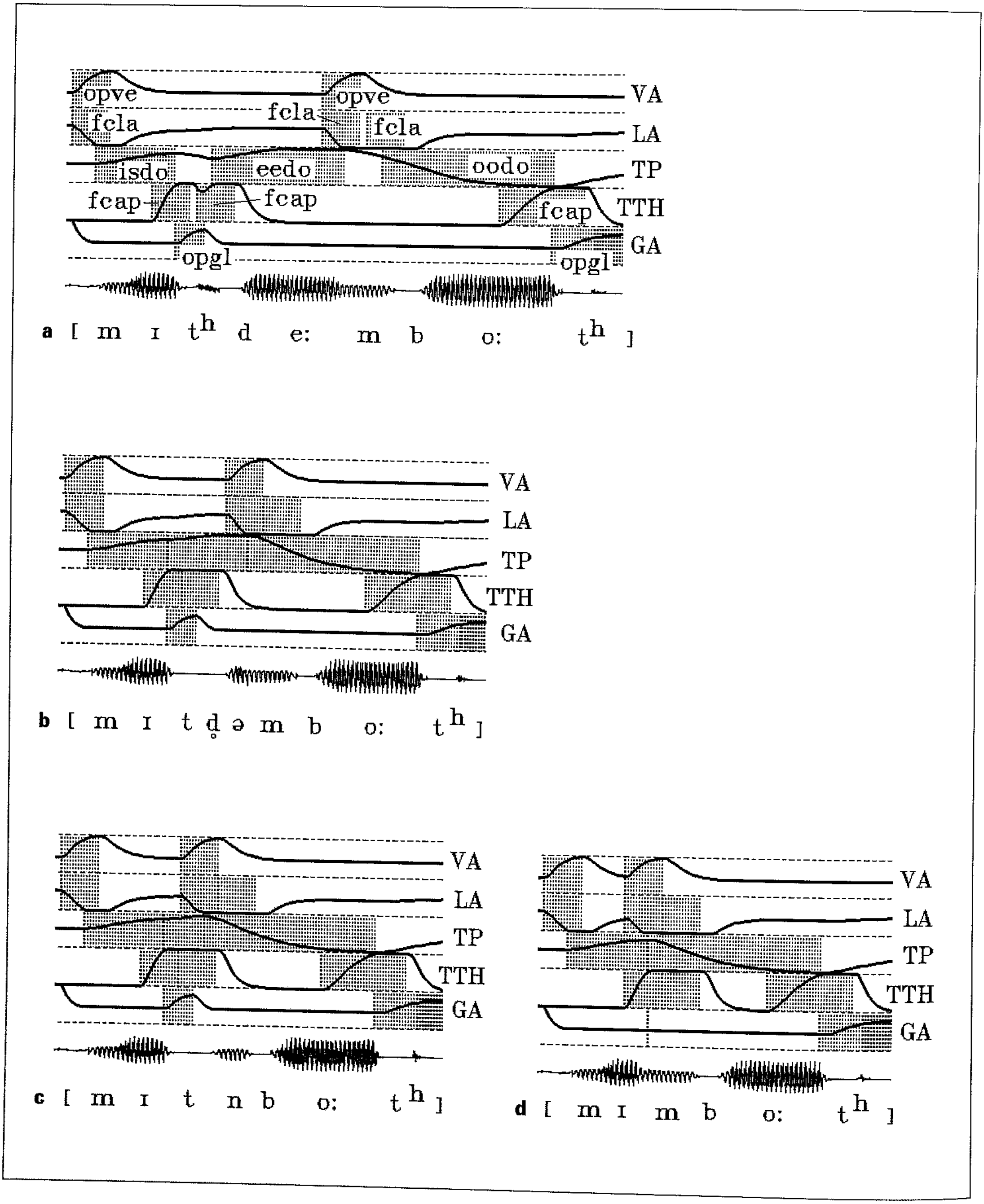

**Fig. 7. a–d** Control parameter time functions (thick lines; for value ranges see fig. 3 and table 1), gestural activation intervals (shaded areas), and the synthetic speech signal for the phrase *mit dem Boot* ('with the boat') and of three reduced forms of this phrase. Only the dorsal /e:/ gesture (eedo) and the glottal opening gesture (opgl) of /mɪt/ are decreased in extension (the release phase value is decreased for these gestures). (Gesture symbols are indicated only in **a**).

**Table 4.** Reduced forms of *mit dem* resulting from several gestural alteration processes

| | 1.0 | 1.1 | 2.0 | 2.1 | 3.0 | 3.1 | 4.0 | 4.1 |
|---|---|---|---|---|---|---|---|---|
| 1 | mɪtʰde:m | – | – | – | – | – | – | – |
| 2 | mɪtd̥e:m | mɪd:e:m | mɪtʰe:m | mɪde:m | – | – | – | – |
| 3 | mɪtd̥əm | mɪd:əm | mɪtʰe:m | mɪde:m | – | – | – | – |
| 4 | mɪt:n | mɪd:n | mɪtʰəm | mɪdəm | – | – | – | – |
| 5 | mɪtn | mɪdn | mɪtn | mɪdn | mɪpm | mɪbm | – | – |
| 6 | mɪp | mɪm | mɪp | mɪm | mɪp | mɪm | – | mɪm |

The empty slots (and total column 4.0) correspond to reduced forms not occurring in German. Vertical dimension: decrease in extension of the dorsal /e:/ gesture. Each line presents a constant degree of extension of this gesture. Horizontal dimension: from column 1.0 to column 2.0: zero to full overlap of the apical full-closing gestures of /mɪt/ and /de:m/. From column 2.0 to column 3.0: change of this blended apical full-closing gesture to a blended labial full-closing gesture (gesture-executing articulator swap from tongue tip to lips). From column 2.0 (or 3.0) to column 4.0: total decrease in extension of this blended apical (or of the swapped blended labial) full-closing gesture. From column x.0 to column x.1 (x = 1, 2, 3, 4): total decrease in extension of the glottal opening gesture of /mɪt/ occurs in addition to the process(es) leading to the forms of column x.0.

hidden by the blended apical full-closing gestures of /mɪt/ and /de:m/. Additionally the blended apical full-closing gestures and the blended labial full-closing gestures of /de:m/ and /bo:t/ overlap. The fully reduced form [mɪmbo:tʰ] results when the dorsal /e:/ gesture and the glottal opening gesture of /mɪt/ are totally decreased in extension (fig. 7d).

In addition to these two concrete processes a third concrete gestural alteration process (consisting of two parts) occurs for this phrase: (3) increase in overlap of the blended apical full-closing gestures of /mɪt/ and /de:m/ and decrease in their extension. It is obvious that the fully reduced form achieved by the two concrete gestural alteration processes stated above (fig. 7d) can also be produced without the blended apical full-closing gestures of /mɪt/ and /de:m/ since their overlap with the blended labial full-closing gestures of /de:m/ and /bo:t/ is synchronous in time.

### *Reduced Forms in the Space of Concrete Gestural Alteration Processes and the Problem of Reduction Hierarchy*

As stated above for our example *mit dem* there are at least three concrete gestural alteration processes which are based on the two types of gestural alteration processes given by Browman and Goldstein [1987, 1989]. The fact that more than one concrete gestural alteration process is involved in the case of our example makes it difficult to establish a one-dimensional reduction hierarchy. The gestural framework in principle provides no constraints against a combination of each degree of realization of one gestural alteration process with each degree of realization of any of the other processes.

Therefore all possible reduced forms stemming from all combinations of all degrees of realization of each process have to be produced. Since every concrete gestural alteration process is described by changing only one gestural descriptor of one distinct gesture, we synthesized groups of stimuli by varying the pertinent gestural descriptor. The transcriptions of these stimuli were arranged in a three-dimensional space where each dimension represents the degree of realization of one gestural alteration process, and then it was collapsed into a two-dimensional table (table 4): the vertical dimension of table 4 represents process 1. The change from column 1 to 2 to 4 (skipping column 3) represents process 3 and the change inside these main colums from x.0 to x.1 (x = 1, 2, 3, or 4) represents process 2. The results of the listening tests given below indicate that the phonemic switches which lead to the reduced forms listed in table 4 are clearly perceptible from the stimuli produced by our synthesizer.

*Additional Gestural Alteration Processes*

Despite the fact that the fully reduced form [mɪmbo:tʰ] can be produced by the above three concrete gestural alteration processes, the form [mɪpmbo:tʰ] [Kohler, 1990, p. 83] cannot be modelled satisfactorily by them (table 4, columns 1, 2, and 4). This form can only be produced if the closure interval of the blended labial full-closing gestures of /de:m/ and /bo:t/ is extended by the closure interval produced by the blended apical full-closing gestures of /mɪt/ and /de:m/.

The principles of articulatory phonology [Browman and Goldstein, 1986, 1987, 1988, 1989] allow a decrease in extension and a change in the phasing of gestures, but gestures (i. e. gestural activation intervals) should not be increased in extension and new gestures should not be introduced in reduction processes aiming for more articulatory simplification. We hypothesize from our modelling experiments that a third type of gestural alteration (or gestural reorganization) is the *gesture-executing articulator swap.* Since no gestural descriptor but the gesture-executing articulator is changed by this type of process, only the articulatory tier is changed, but neither the length nor the location of the gestural activation interval in the time domain is varied.

Therefore a fourth concrete gestural alteration process can be postulated in the case of *mit dem:* (4) the gesture-executing articulator swap of the blended apical full-closing gestures of /mɪt/ and /de:m/ from tongue tip to lips. This fourth process leads to a fourth dimension in the space of concrete gestural alteration processes. Table 4 is not fully extended in order to integrate this fourth dimension since this process requires that process 3 has achieved a total overlap of both apical full-closing gestures (column 2 of table 4). Therefore the effect of this process is integrated as column 3. Segmental changes from column 2 to 3 stem from this gestural swap while changes from column 2 to 4 as well as from 3 to 4 stem from the continuing process of decrease in extension for the blended apical (or even swapped labial) full-closing gestures.

*Listening Tests*

We performed four listening tests in order to evaluate the strength of the perceptual discrimination between adjacent entries listed in table 4. The results of these tests (percentage of responses), the parameters varied for stimulus generation, and the reduced forms which were asked for are given in figure 8 and its legend. The four tests (a–d) together mark a path from the comparatively unreduced form [mɪtd̥e:mbo:tʰ] to the fully reduced form [mɪmbo:tʰ]. For each test seven stimuli were generated by varying only one gestural de-

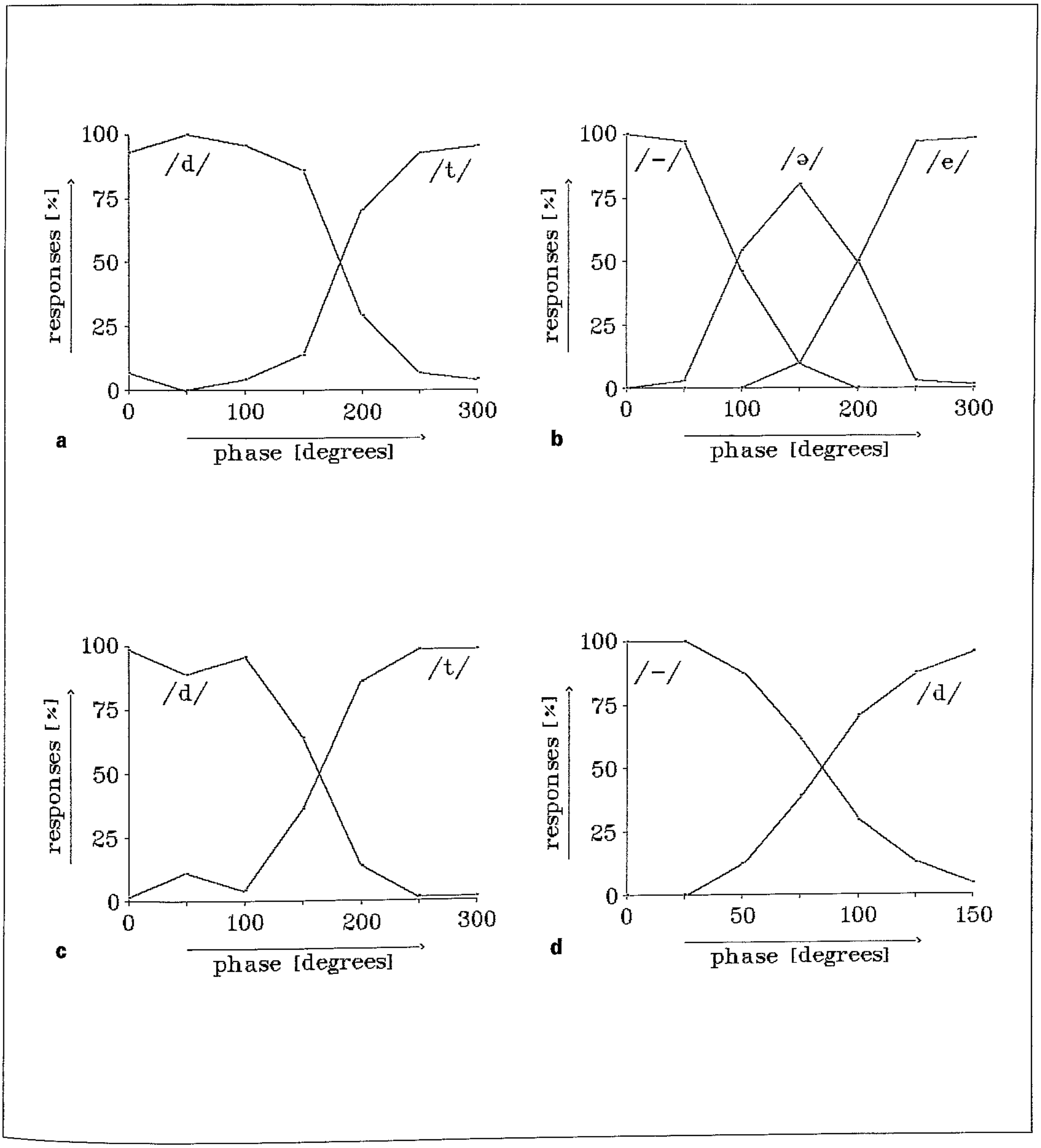

**Fig. 8. a** Percentage of /mɪde:m/ (/d/) versus /mɪte:m/ (/t/) responses as a function of the association phase of the apical full-closing gesture of /de:m/. (0 degrees equals minimal overlap, 300 degrees equals maximal overlap of this gesture and the apical full-closing gesture of /mɪt/. The glottal opening gesture is totally hidden by these blended oral full-closing gestures at 0 degrees.) Transition from line 2, column 1.0 to line 2, column 2.0 in table 4. **b** Percentage of /mɪtm/ (/–/) versus /mɪtəm/ (/ə/) versus /mɪte:m/ (/e:/) responses as a function of the release phase of the dorsal /e:/ gesture. Transition from line 5, column 2.0 to line 3, column 2.0 in table 4. **c** Percentage of /mɪdəm/ (/d/) versus /mɪtəm/ (/t/) responses as a function of the release phase of the glottal opening gesture of /mɪt/. Transition from line 4, column 2.1 to line 4, column 2.0 in table 4. **d** Percentage of /mɪm/ (/–/) versus /mɪd(ə)m/ (/d/) responses as a function of the release phase value of the dorsal /e:/ gesture. Transition from line 6, column 2.1 to line 5, column 2.1 in table 4.

## Appendix

The equation of motion for the damped mass-spring system is

$$m\ddot{x} + b\dot{x} + k\,(x-x_r) = 0 \tag{1}$$

with m = mass, b = damping coefficient, k = stiffnes (spring constant), x = instantaneous position of the mass, $\dot{x}$ = instantaneous velocity of the mass, $\ddot{x}$ = instantaneous acceleration of the mass, $x_r$ = rest position (gestural target). Mass-normalization of this equation leads to a more solution-orientated description of the equation of motion of a linear second-order system:

$$\ddot{x} + \frac{1}{\tau}\dot{x} + \omega_0^2\,(x-x_r) = 0 \tag{2}$$

with $1/\tau = b/m$ and $\omega_0^2 = k/m$. $\tau$ is the relaxation time of the system and $\omega_0/2\pi$ is the eigenfrequency of the undamped system (b=0). Eigenfrequency can be replaced by eigenperiod $T_0$ ($T_0 = 2\pi/\omega_0$). The case of critical damping is defined by

$$\frac{1}{2\tau} = \omega_0 \tag{3}$$

which gives the damping coefficient as a function of mass and stiffness ($b=2\sqrt{km}$). In the case of critical damping, equation 2 can be solved by

$$x\,(t) = x_r + (A+Bt)\,e^{-\omega_0 t} \tag{4}$$

where t is the time and A and B are arbitrary constants needed to fulfill the initial conditions of the system, i. e. $x\,(t=0) = x_0$ and $\dot{x}\,(t=0) = v_0$:

$$A = x_0 - x_r \tag{5a}$$
$$B = v_0 + \omega_0 A. \tag{5b}$$

The case of zero initial velocity ($v_0=0$) leads to the simple equation

$$\gamma\,(\varphi) = (1+\varphi)\,e^{-\varphi} \tag{6}$$

where $\gamma$ is the relative displacement

$$\gamma = \frac{x\,(t) - x_r}{x_0 - x_r} \tag{7}$$

and where $\varphi$ is the phase

$$\varphi = \omega_0 t = 2\pi \frac{t}{T_0}. \tag{8}$$

Table A1 gives some values of the relative displacement as function of phase (see also fig. 5).

In order to avoid discontinuities in the velocity time function of each articulator, gestural time functions are calculated by taking into account the actual initial articulator velocity. But the case of zero initial velocity is a good approximation in many cases, since gestures are mostly activated within the steady-state portion of a gesture or within the steady-state portion of the movement towards the articulator-inherent neutral position.

## References

Browman, C. P.: Consonants and vowels: overlapping gestural organization. Proc. 12th Int. Congr. Phon. Sci., 1991: vol. 1, pp. 379–383.

Browman, C. P.; Goldstein, L.: Towards an articulatory phonology. Phonol. Yb. *3:* 219–252 (1986).

Browman, C. P.; Goldstein, L.: Tiers in articulatory phonology, with some implications for casual speech. Haskins Lab. Status Rep. Speech Res. *92:* 1–30 (1987), and in: Kingston, Beckman, Papers in laboratory phonology. I. Between the grammar and the physics of speech, pp. 341–376 (Cambridge University Press, Cambridge 1990).

Browman, C. P.; Goldstein, L.: Some notes on syllable structure in articulatory phonology. Phonetica *45:* 140–155 (1988).

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

Browman, C. P.; Goldstein, L.: Articulatory phonology: an overview. Phonetica *49:* 155–180 (1992).

Coker, C. H.: Synthesis by rule from articulatory parameters. Proc. 1967 Conf. Speech Commun. Processes, A9, pp. 52–53; reprinted in Flanagan, J. S.; Rabiner, L. R. (eds.): Speech synthesis, pp. 396–399 (Dowden, Hutchingson, & Ross, Stroudsburg 1973).

Edwards, J.; Beckman, M. E.; Fletcher, J.: The articulatory kinematics of final lengthening. J. acoust. Soc. Am. *89:* 369–382 (1991).

Fant, G.: Acoustic theory of speech production (Mouton, The Hague 1960).

Fujimura, O.: Temporal organization of articulatory movements as a multidimensional phrasal structure. Phonetica *38:* 66–83 (1981).

Fujimura, O.: Relative invariance of articulatory movements: an iceberg model; in Perkell, Klatt, Invariance and variability in speech processes, pp. 226–242 (Erlbaum, Hillsdale 1986a).

Fujimura, O.: Evaluating the task dynamics model. J. Phonet. *14:* 105–108 (1986b).

Fujiumura, O.: Towards a model of articulatory control: comments on Browman and Goldstein's paper; in Kingston, Beckman, Papers in laboratory phonology. I. Between the grammar and the physics of speech, pp. 377–381 (Cambridge University Press, Cambridge 1990).

Fujimura, O.: Phonology and phonetics: a syllable-based model of articulatory organization. J. acoust. Soc. Jpn. (E) *13:* 39–48 (1992).

Gay, T.: Mechanisms in the control of speech rate. Phonetica *38:* 148–158 (1981).

Heike, G.: Articulatory measurement and synthesis: methods and preliminary results. Phonetica *36:* 294–301 (1979).

Heike, G.: Articulatory synthesis of German monosyllables; in Brettschneider, Lehmann, Wege zur Universalienforschung. Beiträge zum 60. Geburtstag von Hans-Jakob Seiler, pp. 566–570 (Narr, Tübingen 1980).

Heike, G.: Ein phonologisches Artikulationsmodell des Deutschen; in Feldbusch, Ergebnisse und Aufgaben der Germanistik am Ende des 20. Jahrhunderts. Festschrift für Ludwig Erich Schmitt zum 80. Geburtstag, pp. 677–686 (Olms-Weidmann, Hildesheim 1989).

Ishizaka, K.; Flanagan, J. L.: Synthesis of voiced speech from a two-mass-model of the vocal cords. Bell Syst. tech. J. *51:* 1233–1268 (1972).

Joos, M.: Acoustic phonetics. Lang. Monogr. No. 23. Language *24:* suppl. No. 2 (1948).

Kelly, J. L.; Lochbaum, C. C.: Speech synthesis. Proc. 4th Int. Congr. Acoust., Paper G42 (1962); reprinted in Flanagan, J. S.; Rabiner, L. R. (eds.): Speech synthesis, pp. 127–130 (Dowden, Hutchingson, & Ross, Stroudsburg 1973).

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

Kohler, K. J.: Segmental reduction in connected speech in German: phonological facts and phonetic explanations; in Hardcastle, Marchal, Speech production and speech modelling, pp. 69–92 (Kluwer, Dordrecht 1990).

Kohler, K. J.: The organization of speech production: clues from the study of reduction processes. Proc. 12th Int. Congr. Phon. Sci., 1991a, vol. 1, pp. 102–106.

Kohler, K. J.: The phonetics/phonology issue in the study of articulatory reduction. Phonetica *48:* 180–192 (1991b).

Kröger, B. J.: Three glottal models with different degrees of glottal source-vocal tract interaction. IPKöln-Berichte *16:* 43–58 (1990a).

Kröger, B. J.: A moving noise source and a tube bend in the reflection type line analog. IPKöln-Berichte *16:* 59–69 (1990b).

Kröger, B. J.: Minimal rules for articulatory speech synthesis; in Vandewalle, Boite, Moonen, Oosterlinck, Signal processing. VI. Theories and applications, pp. 331–334 (Elsevier, Amsterdam 1992).

Kröger, B. J.; Heike, G.; Opgen-Rhein, C.; Greisbach, R.; Esser, O.: An investigation of a special type of accentuation in Ripuarian Dialects by computer simulation of speech production. Proc. 12th Int. Congr. Phon. Sci., 1991, vol. 3, pp. 30–33.

Liljencrants, J.: Speech synthesis with a reflection-type line analog; diss. Royal Institute of Technology, Stockholm (1985).

Lindblom, B.: Articulatory activity in vowels. Q. Prog. Status Rep., Speech Transm. Lab., R. Inst. Technol., Stockh., No. 2, pp. 1–5 (1964).

Lindblom, B.: Economy of speech gestures; in MacNeilage, The production of speech, pp. 217–245 (Springer, New York 1983).

Lubker, J.: Articulatory timing and the concept of phase. J. Phonet. *14:* 133–137 (1986).

Mermelstein, P.: Articulatory model for the study of speech production. J. acoust. Soc. Am. *53:* 1070–1082 (1973).

Nittrouer, S.; Munhall, K.; Kelso, J. A. S.; Tuller, B.; Harris, K. S.: Patterns of interarticulator phasing and their relation to linguistic structure. J. acoust. Soc. Am. *84:* 1653–1661 (1988).

Öhman, S. E. G.: Numerical model of coarticulation. J. acoust. Soc. Am. *41:* 310–320 (1967).

Saltzman, E.: Task dynamic coordination of the speech articulators: a preliminary model. Haskins Lab. Status Rep. Speech Res. *84:* 1–18 (1985).

Smith, C. L.; Browman, C. P.: McGowan, R. S.; Kay, B.: Extraction dynamic parameters from speech movement data. Haskins Lab. Status Rep. Speech Res. *105/106:* 107–140 (1991).

Stetson, R. H.: Motor phonetics (North Holland, Amsterdam 1951); reprinted in Kelso, J. A. S.; Munhall, K.G. (eds.): R. H. Stetson's motor phonetics; a retrospective edition (Singular Publ., San Diego 1988).

Stevens, K. N.; House, A. S.: Development of a quantitative description of vowel articulation. J. acoust. Soc. Am. *27:* 484–493 (1955).