3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

A GESTURAL APPROACH FOR CONTROLLING AN ARTICULATORY SPEECH SYNTHESIZER

Bernd J. Kröger

Institut für Phonetik der Universität zu Köln
Köln, Germany

ABSTRACT

Our concept for controlling a speech synthesizer for German is based on articulatory gestures /1,2/. It is distinguished from segmental approaches /3/ by providing a concrete quantitative model of articulatory dynamics. It describes intragestural movement patterns as well as intergestural coordination (gestural phasing).

Keywords: *Speech production, speech synthesis, phonetics*

1. INTRODUCTION

An articulatory synthesis system comprises two main components: (1) A generator for articulator movements, and (2) an articulatory-acoustic model which produces vocal tract shapes and the audio signal. For the first component gestures can be taken as basic units /1/. On one hand gestures are discrete phonological units, describing speech relevant goals like 'labial closure', 'apical closure', 'glottal opening', or 'velic opening'. On the other hand each gesture represents a family of functionally equivalent articulatory movement patterns that are actively controlled with reference to these goals.

2. THE COMPONENTS AND PARAMETERS

A linguistic model generates the gestural score by determining all gestures and the intergestural timing for an utterance (fig. 1). The dynamic model produces continuous control parameter time functions describing the movements of all articulators. The articulatory-acoustic model comprises an articulatory model (based on Heike /4/) which generates vocal tract shapes, and an acoustic model /5,6/ which calculates the audio signal.

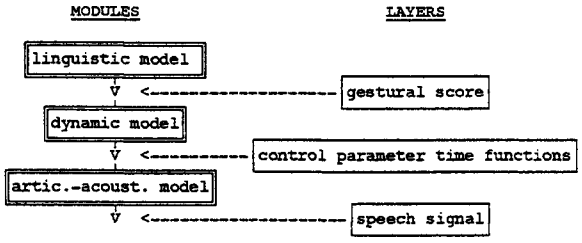
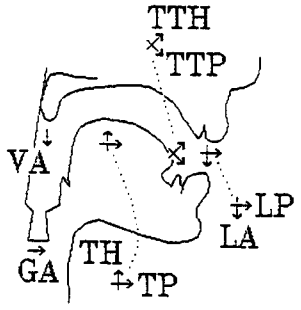


Figure 1 Layers (single lined rectangles) and modules (double lined rectangles) of the articulatory synthesis system

The model operates on the basis of seven articulatory and three phonatory control parameters (fig. 2 and tab. 1). According to the characteristics of our articulatory model the articulatory control parameters directly describe vocal tract shapes, i.e. location and degree of constrictions. This allows us to use a simpler dynamic model than /2/.



articulatory parameters

LA	lip aperture
LP	lip protrusion
TTH	tongue tip height
TTP	tongue tip position
TH	tongue height
TP	tongue position
VA	velic aperture

phonatory parameters

GA	glottal aperture
CT	cord tension
PP	pulmonary pressure

Figure 2 Midsagittal view of the vocal tract shape produced by our articulatory model. The arrows indicate the regions mainly influenced by the control parameters and the directions of major change. (For control parameter definitions see tab. 1)

sym-bol	name	range of values	equivalent articulator position
TH	tongue height	-100 100	lowered (pharyngeal) raised
TP	tongue position	-100 100	back (velar) fronted (palatal)
TTH	tongue tip height	0 100	no elevation occlusion
LA	lip aperture	0 100	closed opened
VA	velic aperture	-100 0 100	strongly raised (closure) raised (closure) lowered (opening)
GA	glottal aperture	-400 0 600	strongly closed (glottal stop) closed (normal phonation) widely opened

Table 1 List of control parameters (selection), range of parameter values, and their equivalent articulator positions. (The extreme values for the control parameters are chosen arbitrarily.)

3. THE CONCEPT 'GESTURE'

Each gesture is active during a definite time interval - its activation interval (fig. 4) - and acts on a distinct articulator performing a movement towards the gestural target. The target represents a characteristic vocal tract constriction. Gestures are phased with respect to each other without reference to an absolute time scale. If no gesture is active for an articulator, this articulator performs a movement towards its inherent neutral position. The neutral position of all articulators defines the production state of a voiced non-nasalized schwa-sound.

4. THE DYNAMIC MODEL FOR GESTURES

The dynamics of each gesture is quantitatively modelled by a critically damped harmonic oscillator /1/. The gestural movement pattern. is an exponential time function asymptotically descending to zero-displacement, i. e. the gestural target (fig. 3). The eigenperiod value of the underlying oscillator defines the time interval needed for reaching a definite relative articulator-target distance. A gesture with a lower eigenperiod value reaches a defined (small) articulator-target distance faster than a gesture with a higher eigenperiod value (see solid-lined and dashed-lined time functions in fig. 3). According to the eigenperiod of a gesture a relative time scale - a phase scale - can be defined for each gesture. Phase values indicate the degree of articulator-target-distance, i.e. the degree to which the gesture has been performed.

We introduce the following rule of thumb for all gestures: 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. That means for example: The longer a gesture is activated above 180 degrees, the more quasi steady state portion of the gesture is produced.

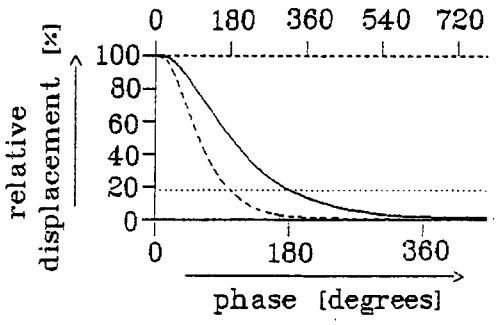


Figure 3 Solid and dashed lines: Time functions of an articulator movement if a gesture is activated for this artiuclator and if the initial articulator velocity is zero (The eigenperiod value is lower in the dashed lined than in the solid lined time function). Abscissa: phase values for both gestures (above for dashed lined gesture, below for solid lined gesture). Ordinate: articulator-target displacement relative to initital displacement. The dotted line indicates clipping.

5. CLIPPING

For plosives or fricatives, the supraglottal constriction-forming gesture must exhibit a temporal interval with a constant degree of constriction. This can only be produced within this quantitative model if the gestural time function 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 3 by the dotted line; in figure 4c it occurs for the dorsal, 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).

For all gestures 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. 3). Therefore the rule of thumb (chap. 4) holds for gestures with unclipped as well as with clipped time functions: The steady state portion of consonantal constriction-forming gestures occurs at phase values above 180 degrees. No consonantal closure for example is produced if the gestural activation ends below 180 degrees for the accompanying full-closing gesture.

6. GESTURAL DESCRIPTORS

In order to describe each gesture and the intergestural timing quantitatively, six gestural descriptors are used: (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. The control parameters define the gesture-performing articulator. For example tongue height is controlled in the case of a dorsal full-closing gesture (tab. 2). Target values (in arbitrary units, see tab. 1) define the direction of each gestural articulator movement; E.g. lip aperture decreases to -20 in the case of labial closing gestures (tab. 2). The clipping value describes the maximal degree of constriction produced by the appertaining gesture, e.g. 0 for a

labial full-closing gesture (tab. 2). The eigenperiod value (in msec) together with the release phase (in degrees) determines the duration of a gesture. The association phase values together with association conventions determine the intergestural coordination (chap. 6).

Phase values define an intrinsic time scale for each gesture (chap. 4). Therefore phase values are more convenient for describing gestural duration and phasing than absolute time values (e.g. in msec). A release phase value indicates the degree to which the appertaining gesture has been executed. For example a release phase value directly indicates wether a consonantal obstruction is produced or to which degree a vowel is reduced or centralized. The phase value description for gestural coordination leads to a local description of articulatory timing. Gestures are not timed with respect to an absolute time scale but with respect to the intrinsic time scales of related gestures (chap. 7).

While the first four descriptors are gesture-inherent in a simple model of speech production (therefore occuring in tab. 2), the other two are context-dependent (occuring in tab. 3). Gestural eigenperiod as well as the gestural target may be context dependent in a more elaborated speech production model which includes e.g. modelling of stress.

sym-bol	name	control para-meter	target value	clipping value	eigen-period
fcla	labial full-closing	LA	-20	0	80
fcap	apical full-closing	TTH	120	100	80
fcdo	dorsal full-closing	TH	120	100	80
ncal	alveolar near-closing	TTH TTP	120 0	96 -	80 80
ncpo	postalveolar near-closing	TTH TTP	120 -50	96 -	80 80
opgl	glottal opening	GA	400	-	80
opve	velic opening	VA	100	-	120
eedo	dorsal /e:/	TH TP	60 90	- -	250 250
oodo	dorsal /o:/	TH TP	40 -100	- -	250 250
osdo	dorsal /o/	TH TP	20 -80	- -	250 250
asdo	dorsal /a/	TH TP	-50 50	- -	250 250

Table 2 List of selected gestures for German, their symbols, and gesture-inherent descriptors: The control parameters involved, values for targets, clipping, and eigenperiod. The dashes indicate absence of clipping.

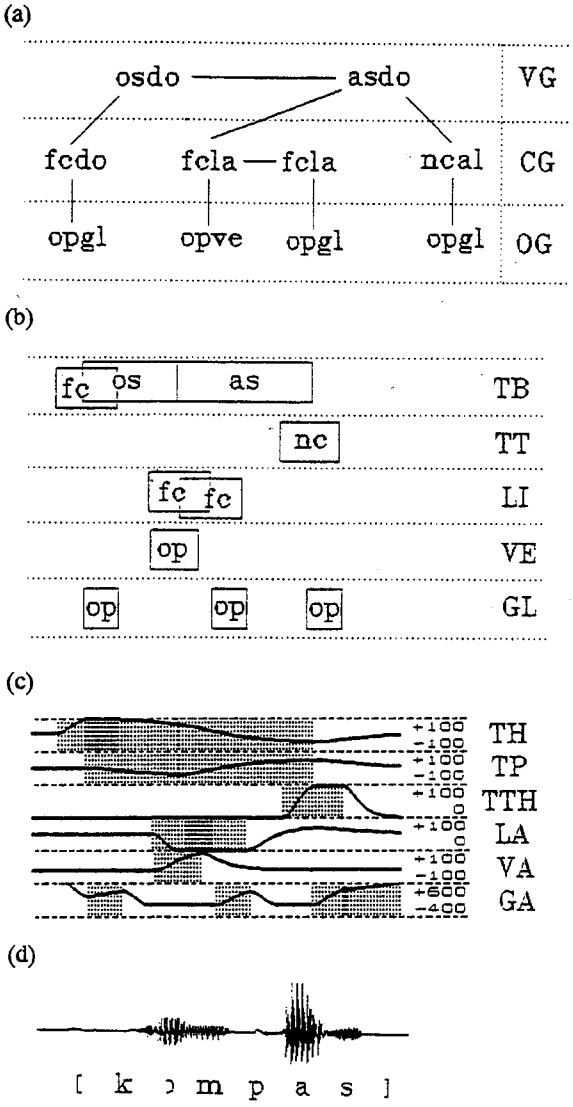


Figure 4 The gestural score of /kompas/ (two different displays), the generated control parameter time functions, and the audio signal. (a) Gestures (for definition of symbols see tab. 2) are ordered in three tiers for vocalic gestures (VG), consonantal gestures (CG) and opening gestures (OG). Association lines indicate which gesture is phased with respect to which other gesture. (b) Gestures are ordered in five tiers according to the gesture-performing articulator: Tongue body (TB), tongue tip (TT), lips (LI), velum (VE) and glottis (GL). The abscissa represents time; Each box represents the gestural activation interval. (c) Control parameter time functions (thick lines) and gestural activation intervals (shaded areas). The last glottal opening gesture is followed by a postphonatory opening gesture. (The definition of control parameter ranges is given in tab. 1.) (d) Oscillogram of the audio signal.

7. TYPES OF GESTURES AND PHASING RULES

Association lines display which gesture is timed or phased with respect to which other gesture (fig. 4a). In order to establish gestural phasing, three types of gestures must be differentiated and arranged in different gestural tiers (fig. 4a): vocalic gestures, consonantal gestures and opening gestures. Vocalic and consonantal gestures produce characteristic vocal tract shapes and constrictions. Examples of consonantal gestures are labial, apical, and dorsal full- and near-closing gestures (tab. 2); Examples of vocalic gestures are /e:/- and /o:/-forming gestures for German long vowels and /o/- and /a/-forming gestures for German short vowels (tab. 2). Only two different opening gestures occur: Velic or glottal gestures.

Both gestures are always associated to consonantal gestures in order to produce nasals or voiceless sounds (tab. 2).
The association phase value of a gesture determines the position of the appertaining gestural phase scale (appertaining activation interval) only with respect to the time instant which is represented by the association line referring to this gesture (That is the association line to the preceding gesture, i.e. to the gesture left or above this gesture; see fig. 4a). In order to define these time instants (represented by association lines) and in order to define which gesture has to be phased with respect to which other gesture (i.e. to define the location of association lines), association conventions are needed. These general conventions are (/1/, pp.11-16 and fig. 4a): (1) Each vocalic gesture is phased with respect to the offset of the preceding vocalic gesture. (2) The first consonantal gesture of a consonant cluster is phased with respect to the onset of the syllable-defining vocalic gesture if the cluster is syllable-initial, and with respect to its offset if the cluster is syllable-final. (3) Non-first consonantal gestures of a consonant cluster are phased with respect to the offset of the preceding consonantal gesture within this cluster. (4) Opening gestures are phased with respect to the offset of the pertinent consonantal gesture.
The association phase values of each gesture together with the above association conventions, determine the intergestural constellation completely. This is illustrated for the German word *Kompaß* ('compass')(fig. 4 and tab. 3).
The conventions given above together with the rules specifying association phase values lead to three main principles for gestural coordination: (1) Vocalic gestures are in an immediate succesion without gaps (Convention (1) plus: Association phase value is zero for vocalic gestures). They act as a 'ground' to consonantal 'figures' /7/. The articulatory movements resulting from these series of vocalic gestures are comparable to the 'vocalic base function' in Fujimura's C/D model /8/ or to the 'vowel component' in Öhman's model /9/. Consequently, consonantal gestures are completely overlapped by vocalic gestures. (2) The consonantal obstruction of a consonant (cluster) coincides with the transient portion of a vowel gesture (Convention (2) plus: Association phase of consonantal gestures is 180 degrees). (3) Consonantal obstructions within a consonant cluster are produced without gaps (Convention (3) plus: Association phase value for consonantal gestures is 180 degrees).
It is remarkable that the release phase value of each vocalic gesture is a syllable-related measure rather than a direct measure of (segmental) vowel duration: As a consequence of principle 1 a vocalic gesture is overlapped by the preceding consonantal gestures. Therefore the release phase value of a vowel depends on the duration and timing characteristics of the gestures forming the consonant cluster which precedes this vowel (e.g. on the number of consonants within this cluster; see tab. 3).

gesture		release phase	association phase
fcdo		400	180
opgl	(1)	180	160
osdo		200	-
fcla	(1)	400	180
opve		200	250
fcla	(2)	400	180
opgl	(2)	180	160
asdo		290	0
ncal		400	180
opgl	(3)	180	160

Table 3 Specification of context-dependent gestural descriptors for /kompas/. The dash for the association phase indicates that this vocalic gesture is not subordinate. The numbers in brackets specify the ordering of equal gestures in time (For graphic display of this gestural score and for the ordering of equal gestures see fig. 4).

8. CONCLUSION

A gestural model for controlling an articulatory synthesizer has been developed for German. This model mainly follows the ideas of Browman and Goldstein /1/ and leads to a favourable description and parameterization of articulatory coordination and articulatory dynamics. More work is needed, especially for the estimation of gestural parameters like eigenperiod, release phase and association phase from real speech samples. Furthermore we plan to model intonation within this framework.

ACKNOWLEDGEMENTS

This work was supported in part by Deutsche Forschungsgemeinschaft DFG grant He 434/21-1 and in part by ESPRIT-BR project Nr. 6975 (SPEECH MAPS).

REFERENCES

/1/: *Browman,C.P.;Goldstein, L.*: Tiers in articulatory phonology, with some implications for casual speech. Haskins Laboratories Status Report on Speech Research SR-92, pp.1-30, 1987
/2/: *Saltzman,E.L.; Munhall,K.G.*: A dynamical approach to gestural patterning in speech production. Ecological psychology, Vol.1, pp.333-382, 1989
/3/: *Kröger,B.J.*: Minimal rules for articulatory speech synthesis. In: J. Vandewalle, R. Boite, et al. (eds.): Signal processing VI: Theories and applications, pp.331-334, 1992 (Amsterdam: Elsevier)
/4/: *Heike,G.*: Articulatory measurement and synthesis. Methods and preliminary results. Phonetica, Vol.36, pp.294-301, 1979
/5/: *Kröger,B.J.*: A moving noise source and a tube bend in the reflection type line analog. IPKöln-Berichte, Vol.16, pp.59-67, 1990
/6/: *Kröger,B.J.*: Three glottal models with different degrees of glottal source - vocal tract interaction. IPKöln-Berichte, Vol.16, pp.43-58, 1990
/7/: *Browman,C.P.*: Consonants and vowels: Overlapping gestural organization. Proc. XIIthInt.Congr.Phon.Sci., Vol.1, pp.379-383, 1991
/8/: *Fujimura,O.*: Phonology and phonetics - A syllable-based model of articulatory organization. J.Acoust.Soc.Jpn.(E), Vol.13, pp.39-48, 1992
/9/: *Öhman,S.E.G.*: Numerical model of coarticulation. J.Acoust.Soc.Am. Vol.41, pp.310-320, 1967