Features and gestures in an articulatory speech production model

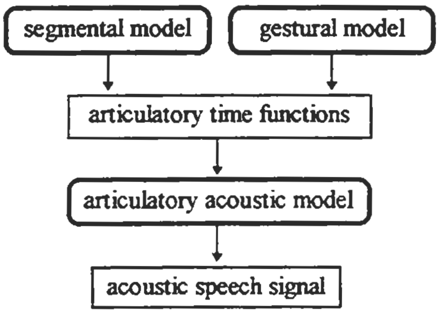
Bernd J. Kröger & Claudia Opgen-Rhein

1. Introduction

A comprehensive computational articulatory speech production model capable of producing any utterance of standard German has been developed. Two different phonological concepts, *gestures* as defined by Browman & Goldstein (1986) and segmental *features* form the basis for two competitive rule components: a gestural and a segmental one. Both produce articulatory time functions, i.e. they produce or define the movements of all articulators (lips, tongue tip, tongue body, velum, and glottis) for an intended utterance (Figure 1). The time functions produced by both components control the same articulatory-acoustic model, which transforms them into a temporal sequence of vocal tract shapes and subsequently into an acoustic speech signal (Kröger 1990a, 1990b; Kröger & Opgen-Rhein 1990).

Features and gestures seem to be the central concepts in speech production modelling. On the one hand they are basic phonological units, i.e. the basic units for the formulation of phonological rules. On the other hand they have an articulatory and also, in the case of features, an acoustic and auditive basis. While a gestural production model is well described (Browman & Goldstein 1990), no complete and computational feature-based articulatory production model can be

Figure 1. The articulatory production model



found in literature. This may be the consequence of problems arising from the fact that many distinctive features cannot easily be interpreted as instructions for articulation.

It is possible to extract basic and indispensable model-articulatory production principles from experience gained during the development of an articulatory speech production model. In our segmental approach these production principles are formalized and defined as *production features*. The production principles as well as the organization of both components are described. Finally the articulatory time functions generated by both components are compared. The parallelism of segmental coarticulation resulting from articulatory underspecification and of gestural coproduction or gestural overlap is explained.

2. The basic production principles in the segmental approach

Input information for the segmental component consists of a phonemic symbol string, which represents a phonetic form, i.e. a broad transcription of the intended utterance. Each phonemic element of this string is converted into a spatial articulatory target configuration which defines the whole or part of the vocal tract shape (Production Principle 1). Target values are assigned to all or a subset of the articulatory (and phonatory) control parameters, i.e. to the parameters which control the positioning of the articulators (Table 1).

Successive non-overlapping time intervals, called production intervals, are defined for each input symbol (as illustrated by the boxes at the top of Figure 2). Each articulatory target has to be reached in its appertaining production interval (Production Principle 2). The time instants at which a target has to be reached are indicated by labels (the vertical dotted lines in Figure 2).

According to contact of an articulator with its counterpart (e.g. lips or vocal folds) or of an articulator with the vocal tract wall (e.g. lower lip with the teeth, tongue tip with the alveolar ridge) a definite and constant degree of vocal tract constriction is produced for an extended time interval (Production Principle 3). This mechanism is important for the production e.g. of plosives or fricatives. Despite the fact that articulators are never at rest during speech production, in the case of contact a defined degree of constriction is reached and kept constant, since the continued articulator movement leads to variation in contact area (cf. Farnetani 1990:111). In our articulatory model, the articulatory control parameters are explicitly defined as the degree and location of vocal tract constrictions and not as the positioning of the centre of mass of each articulator. Therefore we postulate two types of time function (TOT) which can be produced by the model: the unclipped time function (UT) and the clipped time function (CT), the latter

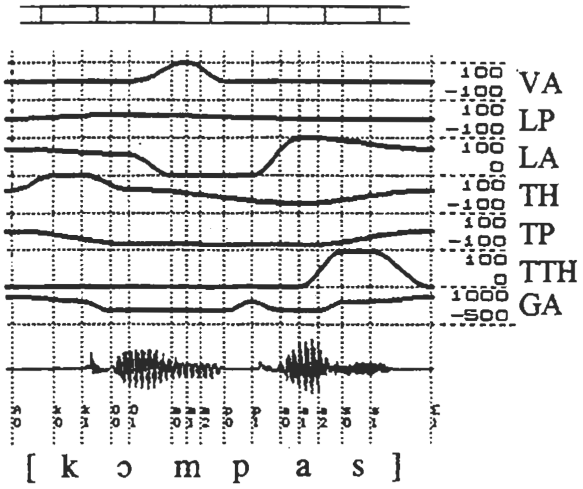
Table 1. (1) Range and meaning of control parameter values. (2) Target information for the generation of the German word "Kompaß" in the segmental approach. Each line represents a control parameter. Each column in (2) represents a phonemic element. For each control parameter and each phonemic element a target value (numbers) and the type of time function (CT: clipped time function, UT: unclipped time function) is given. The empty slots indicate articulatory underspecification.

CONTROL PARAMETER	(1) RANGE and MEANING of values	(2) EXAMPLE:					
		k	ɔ	m	p	a	s
VA	−100 strongly raised (closure)	0	0	100	0	0	0
velic aperture	0 raised (closure)	CT	UT	UT	CT	UT	CT
	100 lowered (opening)						
LP	−100 spread	—	20	—	—	0	—
lip protrusion	0 neutral	—	UT	—	—	UT	—
	100 protruded						
LA	0 closed	—	55	0	0	100	—
lip aperture	100 open	—	UT	CT	CT	UT	—
TH	−100 lowered (pharyngeal)	100	25	—	—	-50	—
tongue (body)	0 neutral	CT	UT	—	—	UT	—
height	100 raised						
TP	−100 back (velar)	—	-70	—	—	-80	—
tongue (body)	0 neutral	—	UT	—	—	UT	—
position	100 fronted (palatal)						
TTH	0 no elevation	—	0	—	—	0	96
tongue tip height	100 occlusion	—	UT	—	—	UT	CT
GA	−500 strongly closed (glottal stop)	400	10	10	400	10	400
glottal aperture	0 closed (normal phonation)	UT	CT	CT	UT	CT	CT
	1000 wide open						

resulting from the contact mechanism described above. The type of time function must be specified for each phonemic element and for each articulator (Table 1). For an unclipped time function, one label is assigned to the middle of the production interval (Figure 2). For a clipped time function the control parameter value is kept constant for nearly the full production interval by setting one label near each end of the production interval. Since even for one input element different articulators can exhibit different types of time function, up to three labels may occur at one production interval.)

Despite some shortcomings of a segmental target theory (MacNeilage 1980) our work shows that a comprehensive production model can be developed on the basis of this theory if the clipping mechanism is taken into account. Additionally, great care must be taken to minimize the set of control parameters for which target values must be specified for a given phonemic element. *Articulatory underspecification* has to be maximal in order to derive sufficient allophonic variability (Production Principle 4). The unspecified articulators for each phonemic element are indicated by empty slots in Table 1. For example in the case of /p/ only the lips have to be controlled by defining a spatial lip target while the other tract-shaping articulators (tongue tip and tongue body) are free. To be more precise

Figure 2. Selected control parameter time functions (articulatory time functions) and oscillogram of the synthetic speech signal generated by the segmental component (for the definition of the control parameters see Table 1). The boxes at the top indicate segmental production intervals.



only a target for lip aperture has to be fixed in order to allow additionally any desired degree of lip protrusion. In the case of /k/ only a target for tongue body height has to be fixed. Therefore the tongue body position is defined by context, e.g. a palatal constriction is produced for /gi/ whereas the constriction is velar for /gu/. Therefore underspecification as defined in this framework is the source for coarticulation (see also §3 below). But our model is not so detailed, that it can deal with cross language aspects of coarticulation, e.g. as investigated by McAllister & Engstrand (1991).

These four basic production principles are sufficient to establish an articulatory speech production model. But we have not found a straightforward way to relate distinctive features directly to these production principles. Although distinctive features were defined in terms of speech production by Chomsky & Halle (1968) it is very difficult to interpret all distinctive features as instructions to articulators. Therefore we define a set of *production features* which are directly related to the production principles stated above. A definite discrete value must be specified for each production feature: (1) The feature 'articulatory underspecification' (AUS) can be specified as 'full specification' (AUS=FS) or as 'underspecification' (AUS=US). (2) In the case of underspecification only one tract-shaping articulator has to be specified by evaluating the production feature 'constriction forming articulator' (CFA) as 'lips' (CFA=LI), as 'tongue tip' (CFA=TT), as 'tongue body' (CFA=TB), or as 'glottis' (CFA=GL). (3) The 'type of time function' (TOT) has to be specified as 'clipped time function' (TOT=CT) or as 'unclipped time function' (TOT=UT) for each phonemic element and for each articulator. (4) Target values have to be fixed for the selected control parameters for each phonemic element (e.g. degree of constriction in the case of plosives, or tongue height and position in the case of vowels).

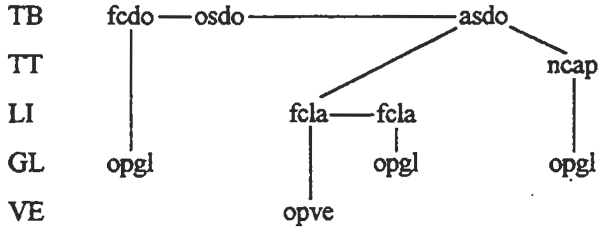
In order to avoid distinctive features it is possible to establish a phoneme conversion list for converting each phonemic element directly into production features. However in such a list several classes of phonemic elements appear which show an identical subset of production features all specified by an identical pattern of feature values. For example: (1) Different patterns of production feature values which determine the state of the glottis (TOT and target for glottal aperture) occur for the two classes: voiced and unvoiced sounds. (2) Different patterns of production features which determine the state of the velum (TOT and target for velic aperture) occur for the three classes: nasals, obstruents, and nonnasal sonorants. (3) Different feature value patterns for all production features except CFA occur for classes of consonants with different manners of articulation (e.g. nasals, fricatives, plosives). Therefore a preanalysis of the phonemic symbol string with regard to laryngeal state, manner of articulation, and place of articulation leads to fewer repetitions of identical patterns of specified production features. A similar division into three parts can be found in feature geometry approaches (Clements 1985:248, McCarthy 1988) and the grouping of features is also a basic idea in dependency phonology (Anderson & Ewen 1987).

3. Comparison of articulatory time functions generated in the segmental and the gestural approach

Our gestural component was developed closely following the concept devised by Browman and Goldstein (1990). A detailed description of the model is given in Kröger (1993). Gestures are filed in articulatory tiers according to the end-articulator which executes them (Figure 3). Phasing lines indicate which gesture is phased or timed with regard to which other gesture. The shaded boxes in the time function areas (Figure 4) mark the activation interval of each gesture, i.e. the time interval in which it is active. In contrast to the linearly ordered non-overlapping segmental production intervals, gestures are organized on more than one tier and their activation intervals can overlap in the time domain.

The comparison of the articulatory time functions generated by both components (Figures 2 and 4) reveals a great similarity. On the one hand this is not surprising since both models produce the same word "Kompaß", on the other hand it is remarkable since both components were developed on the basis of completely different concepts: segmental features versus gestures. A comparison of some characteristics of both approaches elucidates the background for this similarity. (1) Both components produce only goal-directed movements. In the seg-

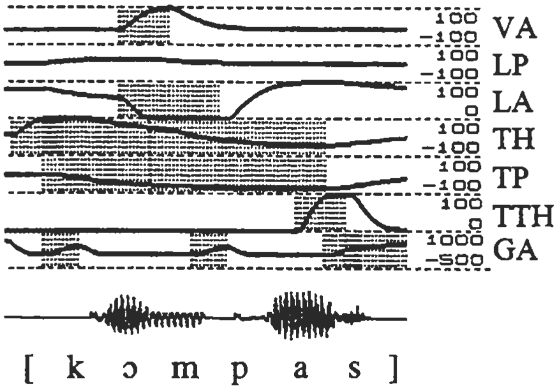
***Figure** 3. Gestures of the word "Kompaß" labelled by their symbols and filed in articulatory tiers according to the end-articulator which executes each gesture. Phasing relations are indicated by phasing lines. Articulatory tiers: TB: tongue body, TT: tongue tip, LI: lips, GL: glottis, VE: velum. Gestures: fcdo: dorsal full-closing gesture, osdo: dorsal short /o/-forming gesture, asdo: dorsal short /a/-forming gesture, ncap: apical near-closing gesture, fcla: labial full-closing gesture, opgl: glottal opening gesture, opve: velic opening gesture.*



mental approach all time functions exhibit smooth and short transitions from target to target. In the gestural approach gestures are directly defined as goal-directed articulatory movements. (2) The clipping mechanism is implemented in both components. (3) The principle of articulatory underspecification leads to coarticulation in the segmental model. This segmental coarticulation is equivalent to gestural coproduction or gestural overlap. For example the tongue body movement during the production of labial consonants in the segmental model (see the tongue body movement for /m/ and /p/ in Figure 2) resembles the tongue body movement during the labial gestures in the gestural model (see the tongue body movement during the activation of the short /a/-forming gesture in Figure 4) since the activation of vocalic gestures always begins during the activation of the tract-shaping gesture(s) of the preceding consonant (cluster). It is an overall gestural production principle that consonantal gestures overlap vocalic gestures (Browman 1991). Another example for segmental coarticulation and gestural coproduction is the variation of lip protrusion and tongue position during the dorsal closure of /k/, or of lip protrusion during the labial closure of /m/ and /p/ (Figures 2 and 4).

Similarity of the time functions produced by both models ends when it comes to modelling phenomena like articulatory undershoot variation (due to stress) or like the occurrence of different kinds of discrete segmental changes (due to reduction, e.g. Kröger 1993). Modelling these phenomena appears to be out of the scope of any segmental rule system. But in the gestural approach target undershoot variation according to stress or speech rate is a straightforward consequence of limitations on gestural stiffness and of limitations on the length of gestural activation intervals. Moreover different kinds of discrete segmental changes implying reduction which have to be incorporated each as a segmental rule in a segmental approach (e.g. vowel reduction to schwa, schwa-elision, assimilation of place and voice, and consonant gemination and elision) can be reproduced

Figure 4. Selected control parameter time functions and oscillogram of the synthetic speech signal generated by the gestural component (for the definition of the control parameters see Table 1). The shaded boxes mark the activation intervals for each gesture (see Figure 3 for the name of each gesture). A detailed description of the generation of articulator time functions is given in Kröger (1993).



easily and straightforwardly in a gestural approach: only two underlying continuous gestural alteration processes, the increase in gestural overlap and the decrease in gestural extent, are responsible for these different kinds of discrete segmental changes.

4. Conclusion

We have compared the articulatory time functions produced by a segmental and a gestural speech production component and stressed the importance of clipping due to contact and of articulatory underspecification in segmental concepts (for a discussion of underspecification in phonetics see Boyce, Krakow et al. 1991, and Keating 1988). The parallelism of segmental coarticulation and gestural coproduction has been elucidated (see also Fowler 1980).

While gestures are by definition closely related to speech production this does not hold for distinctive features. During the development of a control component for our articulatory synthesizer we found no straightforward way of basing a segmental rule component on Chomsky & Halle's (1968) distinctive features. This results in part from the non-orthogonality of distinctive features with respect to the system of speech articulators. "There are many features that cannot be interpreted as instructions to the articulators without knowing what the values of the other features are." (Ladefoged 1992:175-176). Additionally it should be emphasized that our system of production features is not introduced in order to relate a gestural to a segmental approach; it originates exclusively from renaming the underlying segmental production principles in terms of features, which must be specified in a definite way for representing or producing speech sounds.

While Kröger (1993) focuses on advantages of a gestural in comparison to a segmental description of speech, this paper has tried to elucidate some similarities of both approaches. It remains to be shown by future research that the concept of gestures as developed in the framework of articulatory phonology (Browman and Goldstein 1986) is able to model as great a variety of phonological processes as are described by segmental phonological rules using distinctive features. It should become possible to formulate the phonology of a language — as has successfully been done by segmental feature-based generative approaches (cf. Wurzel 1970 for German) — in terms of a gestural framework.

References

Anderson, J.M. & C.J. Ewen. 1987. *Principles of dependency phonology.* Cambridge: Cambridge University Press.

Boyce, S.E., R.A. Krakow, & F. Bell-Berti. 1991. Phonological underspecification and speech motor organization. *Haskins Laboratories Status Report on Speech Research* SR-105/106:141-152.

Browman, C.P. 1991. Consonants and vowels: Overlapping gestural organization. *Proceedings of the XIIth International Congress of Phonetic Sciences.* Vol. 1: 379-383.

Browman, C.P. & L. Goldstein. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3:219-252.

Browman, C.P. & L. Goldstein. 1990. Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M.E. Beckman (eds.), *Papers in laboratory phonology I, Between the grammar and physics of speech.* Cambridge: Cambridge University Press, 341-376.

Chomsky, N. & M. Halle. 1968. *The sound pattern of English.* New York: Harper & Row.

Clements, G.N. 1985. The geometry of phonological features. *Phonology Yearbook* 2:225-252.

Farnetani, E. 1990. V-C-V lingual coarticulation and its spatiotemporal domain. In W.J. Hardcastle & A. Marchal (eds.), *Speech production and speech modelling.* Dordrecht: Kluwer, 93-130.

Fowler, C.A. 1980. Coarticulation and theories of extrinsic timing. *Journal of Phonetics* 8:113-133.

Keating, P.A. 1988. Underspecification in phonetics. *Phonology* 5:275-292.

Kröger, B.J. 1990a. A moving noise source and a tube bend in the reflection type line analog. *IPKöln-Berichte* 16:59-67.

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

Kröger, B.J. 1993. A gestural production model and its application to reduction in German. *Phonetica* 50:213-233.

Kröger, B.J. & C. Opgen-Rhein. 1990. Das physiologisch - akustische Modell AKUSYN des artikulatorischen Sprachsynthesesystems ASSCO. *IPKöln-Berichte* 16:69-118.

Ladefoged, P. 1992. The many interfaces between phonetics and phonology. In W.U. Dressler, H.C. Luschützky, O.E. Pfeiffer, & J.R. Rennison (eds.), *Phonologica 1988. Proceedings of the 6th International Phonology Meeting.* Cambridge: Cambridge University Press, 165-179.

MacNeilage, P.F. 1980. Distinctive properties of speech motor control. In G.E. Stelmach & J. Requin (eds.), *Tutorials in Motor Behavior.* Amsterdam: North-Holland, 607-621.

McAllister, R. & O. Engstrand 1991. Some cross language aspects of co-articulation. *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm* (PERILUS) 14:7-10. University of Stockholm.

McCarthy, J.J. 1988. Feature geometry and dependency: A review. *Phonetica* 43:84-108.

Wurzel, W.U. 1970. Studien zur deutschen Lautstruktur. *Studia Grammatica* 8. Berlin: Akademie Verlag.