SIGNAL PROCESSING VI: Theories and Applications
J. Vandewalle, R. Boite, M. Moonen, A. Oosterlinck (eds.)

MINIMAL RULES FOR ARTICULATORY SPEECH SYNTHESIS

Bernd J. KRÖGER

Institut für Phonetik der Universität zu Köln
Greinstr. 2, D-5000 Köln 41

A concept of minimal segmental rules for the control of an articulatory speech synthesizer is presented. Production features are introduced as a vehicle to convert distinctive features in production instructions. The model produces articulatory movements which show a good modelling of coarticulation and lead to intelligible synthetic speech.

1. Introduction

The aim of articulatory speech synthesis is to model the human speech production mechanisms as closely as possible. This may be the key to overcome the quality limitations, existing for acoustics-based speech synthesizers. Additionally, articulatory synthesizers can be used as a tool for developing an articulatory phonology (BROWMAN and GOLDSTEIN 1987). Articulation-based rule programs (SONDHI and SCHROETER 1987; MEYER, WILHELMS et al. 1989) are rare because of their complexity and because of their great computational cost.

2. The modules and levels of the production model

Our production model converts a quasi phonemic symbol string into an acoustic speech signal (fig. 1). Every symbol is changed into a vector of distinctive features and converted into production features leading to the target grid (chap. 4). The dynamic model produces continuous articulatory and phonatory control parameter trajectories (The definition of the control parameters is given in fig. 2 and tab. 1). The articulatory control parameters pass the articulatory model, which converts them into midsagittal vocal tract shapes. The geometric-acoustic transformation transforms these shapes into area functions of equally long cylindrical tube segments. The acoustic model produces the acoustic speech signal.

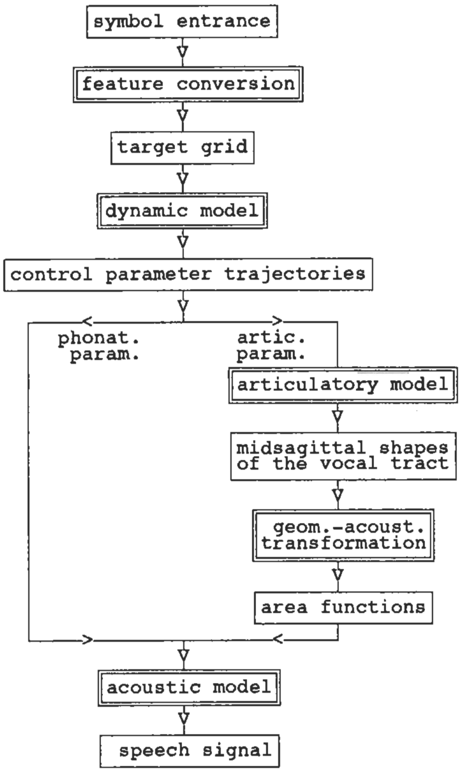
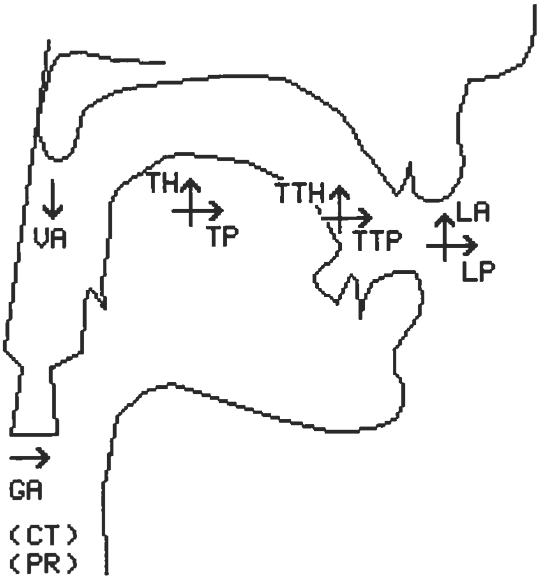


Figure 1 The production model: levels (single lined rectangles) and modules (double lined rectangles)



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
PR	lung pressure

Figure 2 Midsagittal shape of the vocal tract produced by the articulatory model and a list of the control parameters

CONTROL PARAMETER	RANGE of values MEANING of values		
VA velic aperture	-100 strong closure	0 closure	100 wide opening
LA lip aperture		0 closure	100 wide opening
TH tongue height	-100 low position	0 neutral position	100 high position
TP tongue position	-100 back position	0 neutral position	100 front position
TTH tongue tip height		0 neutral position	100 high position
TTP tongue tip position	-100 postalv. position	0 alveolar position	100 dental position
GA glottal aperture	-500 strong closure	0 closure (phonation)	1000 wide opening

Table 1 Range of (relative) values and meaning of extreme and mean values of selected control parameters.

CONTROL PARAMETER	EXAMPLE ?	a	p	S	I	k	@	n
VA velic aperture	- -	0 FM	0 LM	0 LM	0 FM	0 LM	0 FM	100 FM
LA lip aperture	- -	100 FM	0 LM	- -	40 FM	- -	70 FM	- -
TH tongue height	- -	-50 FM	- -	- -	70 FM	100 LM	20 FM	- -
TP tongue position	- -	-80 FM	- -	- -	70 FM	- -	0 FM	- -
TTH tongue tip height	- -	0 FM	- -	96 LM	0 FM	- -	0 FM	100 LM
TTP tongue tip position	- -	0 FM	- -	-100 LM	0 FM	0 LM	0 FM	- -
GA glottal aperture	-400 LM	10 LM	400 FM	400 FM	10 LM	400 FM	10 LM	10 LM

Table 2 Target values and type of articulator movement (TAM=LM: limited articulator movement; TAM=FM: free articulator movement) for /?apSIk@n/ ("to send off"). /?/ ≙ glottal stop; /S/ ≙ postalveolar voiceless fricative; /@/ ≙ schwa-sound. For control parameters see fig. 2 and tab. 1. Limited articulator movement (LM) leads to two labels, one at each end of the production interval, and free articulator movement (FM) leads to one label, mostly in the center of the production interval (see fig. 3). Passive articulators are indicated by dashes.

3. The acoustic model

The acoustic model which comprises the vocal tract model, a reflection type line analog (KRÖGER 1990a) and a self-oscillating glottal model (KRÖGER 1990b) reproduces the main acoustic, physiologic, and aerodynamic characteristics of speech production. The vocal cord vibration is controlled by lung pressure, cord tension, and glottal aperture. The location and amplitude of the friction noise is calculated from the oral constriction area and from the volume flow of the air streaming through this constriction. A direct control of the noise amplitude and location is not necessary. But this model requires an excellent cooperation of lung pressure, glottal aperture and oral constriction degree to get realistic magnitudes of airflow necessary for the insertion of friction noise.

4. Feature conversion and production features

The production model is intended to be connected to a segmental phonological component (e.g. Wurzel, 1970). A quasi phonemic symbol string serves as input for the rule component (feature conversion and dynamic model) which produces a set of continuous control parameter trajectories. In a first step, the quasi phonemic input symbols are transformed into the target grid. This step is called feature conversion, since every input symbol can be seen as a vector of (segmental phonological) distinctive features and since the target grid is only one possible or concrete realization of underlying *production features*.

The labels of all input symbols form the target grid. They define the time instants at which the targets (spatial goals of articulatory movements) must be reached by the articulators (tab. 2 and fig. 3). A time interval, called *production interval*, is defined for every input symbol. The labels belonging to an input symbol occur in its production interval.

The production feature *articulatory underspecification* (AUS) differentiates the production of vowels and glides (distinctive feature [+voc]) where targets are defined for all articulators (AUS=FS, full specification) from the production of liquids, nasals, and obstruents (distinctive feature [-voc]) where targets are defined only for the constriction forming articulator (AUS=US underspecification). So we have to differentiate between active articulators (defined target) and passive articulators for every sound. In the case of underspecification, the production feature *constriction forming articulator* (CFA) determines an active articulator (e.g. lips in the case of bilabial plosives).

Articulatory underspecification leads to a high degree of coarticulatory freedom and produces the allophonic variation for the production of phonemes in different contexts. For example in the case of the bilabial consonant /p/ (fig. 3), the tongue tip and the tongue body are passive articulators. The transition of tongue tip and tongue body control parameter trajectories is not influenced by this consonant but by the surrounding sounds.

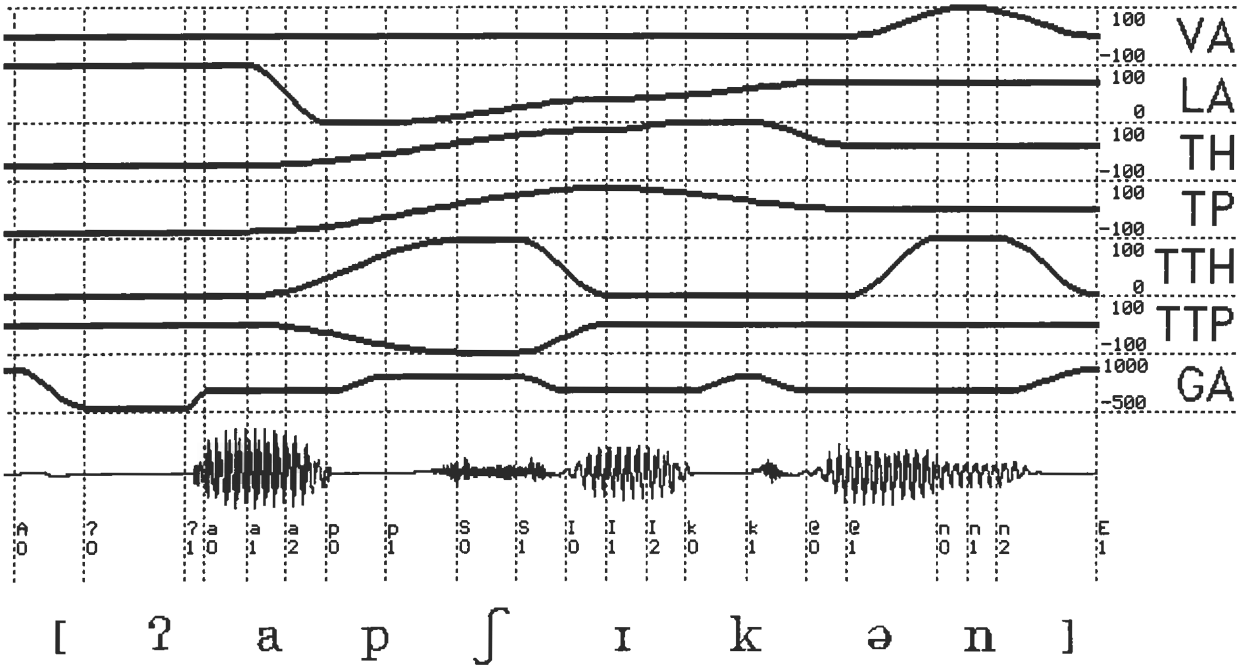


Figure 3 Control parameter trajectories (thick lines), oscillogram, and phonetic transcription of the synthetic speech signal for the input symbol sequence /?apSIk@n/ ("to send off"). The horizontal dashed lines separate different control parameter areas. The vertical dashed lines indicate time instants (labels) when target values given by the target grid (tab. 2) are reached. One to three labels are generated for each input symbol, fixing articulatory and phonatory control parameters. Passive articulators of a quasi phoneme (see tab. 2) are not affected by its labels. Bottom of each label: input symbol and current label number.

The production feature *type of articulator movement* (TAM) differentiates between limited (articulator) movements (TAM=LM) and free (articulator) movements (TAM=FM). From a physiological viewpoint, the limited movement results (1) from strong contact of the articulator with the palate (or teeth or upper lip) which occurs at an oral constriction (distinctive feature [+cons]), (2) from the strong contact of the velum with the pharynx wall, which occurrs in obstruents to stop air leakage through the nasal tract (distinctive feature [-son]), and (3) from the contact of the vocal cords to ensure phonation (distinctive feature [+voice]). The term "limited movement" is chosen, since this type of movement can be understood as a normal target-directed movement, which is stopped or limited by contact or collision of the articulator with rigid walls or with its counterpart. This limited articulator movement is realised in our production model by holding the control parameter constant throughout the whole production interval. Two labels are set up, one at each end of the production interval. Free articulator movements are produced if the articulator reaches the defined target without restraints by contacts. Free articulator movements are realized in our model by setting up only one label in the middle of the production interval.

The resulting number of labels representing an input symbol in the target grid belongs to the TAM-specifications for all articulators involved in the production of this sound. In the case of a vowel surrounded by two voiceless consonants (e.g. /I/ in fig. 3), label 0 and label 2 fix the glottal aperture (TAM=LM) which ensures voicing and label 1 fixes the vocal tract articulators (TAM=FM). Even in the case of vowels, no steady state portion of the vocal tract shape is produced. The targets of the vocal tract shape forming articulators are reached only in the middle of the production interval.

The last group of production features comprises the target values themselves. For every sound, every active articulator needs a target to define the goal-directed movement needed for the production of this sound.

There are some parallels and some differences between distinctive features and production features. While the production features AUS and TAM are binary features and the production feature CFA is a n-ary feature (4 possible articulators), the targets are production features using a continuous scaling. Like the distinctive features, the production features AUS and CFA are only related to the segment itself. But the production feature TAM and the targets must be specified for every segment and also for every active articulator. And it must be emphasized that no one-to-one relation exists between distinctive features and production features. Thus the production feature TAM for different articulators is specified by different distinctive features ([±cons], [±son], and [±voice]) while the distinctive feature [±voice] also determines the target for the glottal articulator. The good intelligibility of our synthesis system leads us to believe that the kind of conversion described here is a realistic way to convert distinctive features in phonetic production instructions. Attempts at a more direct conversion have failed.

5. Conclusions

A concept of minimal rules for articulatory speech synthesis is introduced. The great amount of allophonic variation is automatically produced by articulatory underspecification: Passive articulators are coarticulated widely by surrounding sounds. The segmental rules described here are not context dependent. A production rule for a quasi phoneme need not be changed according to surrounding sounds.

The production model given here is a segmental phonological or linear phonological model. Successive or linearly ordered production intervals are defined for the linearly ordered input symbols. Segmental models have shortcomings in modelling suprasegmental phenomena (accentuation, intonation) and in modelling reduction phenomena. In future work a non-linear gestural phonological production model (BROWMAN and GOLDSTEIN 1987) will be implemented on the basis of the model, described above.

References

Browman, C. P., L. Goldstein (1987): "Tiers in articulatory phonology, with some implications for casual speech", *Haskins Laboratories Status Report on Speech Research* ***SR-92***, 1-30.

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.

Meyer, P., R. Wilhelms, H. W. Strube (1989): "A quasiarticulatory speech synthesizer for German language running in real time", *Journal of the Acoustical Society of America* ***86***, 523-539.

Sondhi, M. M., J. Schroeter (1987): "A hybrid time-frequency domain articulatory speech synthesizer", *IEEE Transactions on Acoustics, Speech, and Signal Processing* ***ASSP-35***, 955-967.

Wurzel, W. U. (1970): "Studien zur deutschen Lautstruktur", *studia grammatica* ***VIII***, Berlin: Akademie-Verlag.