A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis

Bernd J. Kröger[1] and Peter Birkholz[2]

[1] Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital Aachen and Aachen University, 52074 Aachen, Germany
[2] Institute for Computer Science, University of Rostock, 18059 Rostock, Germany
bkroeger@ukaachen.de
piet@informatik.uni-rostock.de

Abstract. An articulatory speech synthesizer comprising a three-dimensional vocal tract model and a gesture-based concept for control of articulatory movements is introduced and discussed in this paper. A modular learning concept based on speech perception is outlined for the creation of gestural control rules. The learning concept includes on sensory feedback information for articulatory states produced by the model itself, and auditory and visual information of speech items produced by external speakers. The complete model (control module and synthesizer) is capable of producing high-quality synthetic speech signals and introduces a scheme for the natural speech production and speech perception processes.

Keywords: articulation, speech synthesis, speech production, gesture.

1 Introduction

Within the field of cross-modal analysis of verbal as well as nonverbal communication it is feasible to use approaches capable of modeling speech perception and speech production as close as possible to real biological processes. But neural models of speech perception and speech production are rare (e.g. Guenther 2006, Guenther et al. 2006, Kröger et al. 2006b, Kröger et al. 2007). A general approach for modeling speech production and perception is introduced in this paper. We describe a comprehensive, computer-implemented gesture-based model for speech movement control and outline learning procedures for the optimization of control rules within this model. The learning procedures are based on auditory as well as on visual speech perception.

2 Three-Dimensional Model of the Vocal Tract and Acoustic Simulation

From an acoustic point of view, the vocal and nasal tracts constitute lossy tubes of non-uniform cross-sectional area. With regard to a computer simulation of acoustic

A. Esposito et al. (Eds.): Verbal and Nonverbal Commun. Behaviours, LNAI 4775, pp. 174–189, 2007.

propagation within these tubes, as it is needed for an articulatory speech synthesizer, the following simplifying assumptions can be made to keep the numerical calulations tractable (Flanagan 1965): (1) Sound propagates as plane waves, i.e., we assume one-dimensional wave propagation. (2) We neglect tube bendings, especially between the mouth and the pharynx. (3) The tubes are represented in terms of incremental sections of right, elliptical geometry.

With these assumtions, the vocal system can be represented as a branched tube system consisting of short abbutting tube sections as illustrated schematically in Fig. 1. This tube model, in turn, can be represented by an area function, i.e., the cross-sectional areas along the centerline of the tube branches. In our synthesizer, this area function is generated by an articulatory model of the vocal tract and a model of the vocal folds that will be described in Section 2.1. The simulation of acoustic propagation within the tube model and the generation of the synthetic speech signal will be briefly discussed in Section 2.2.

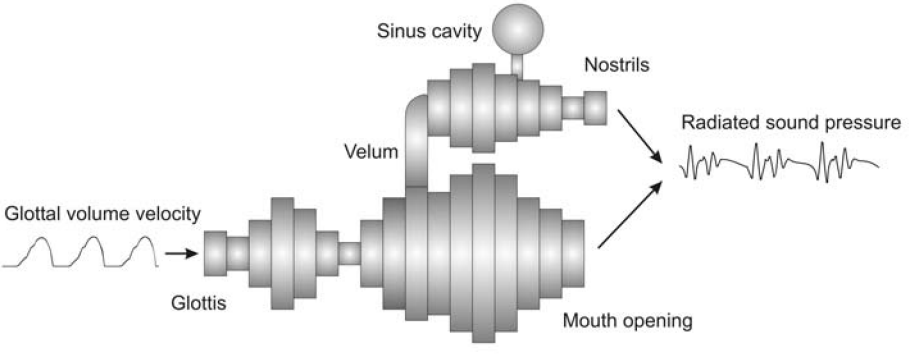

Fig. 1. Schematic representation of the tube model for the vocal tract and the nasal tract

2.1 Vocal Tract Model

The vocal tract is modeled in terms of a number of wireframe meshes that represent the surfaces of the articulators and the vocal tract walls. Fig. 2 gives an overview of the meshes. Two of them, the upper cover and the lower cover, represent the superior-posterior and the inferior-anterior walls of the vocal tract, respectively. The upper cover comprises sub-surfaces for the posterior wall of the larynx and the pharynx, the velum, and the hard palate. The lower cover consists of sub-surfaces for the anterior wall of the larynx and pharynx, and the mandible. The tongue, the upper and lower teeth, the upper and lower lip, the epiglottis, and the uvula are all represented by individual surfaces. Fig. 5 (b) shows a 3D rendering of the vocal tract for the vowel [i:].

All of the surfaces are defined in the fixed reference frame of the hard palate. The teeth, mandible, palate, uvula, and epiglottis are considered as rigid structures, while the velum, the tongue, the lips, and the laryngeal and pharyngeal walls are deformable. The shape of the rigid structures has been adapted to an adult male speaker by

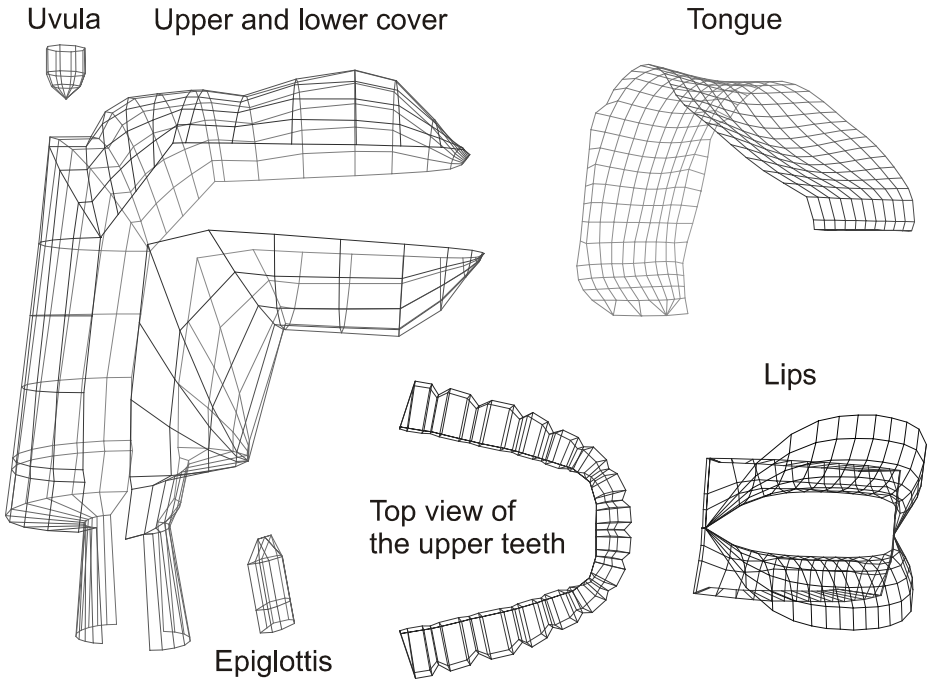

Fig. 2. Wireframe representation of the surfaces defining the vocal tract geometry

means of volumetric Magnetic Resonance Imaging (MRI) data of the vocal tract and Computer Tomography scans of plaster casts of the jaw and teeth, and therefore represent his anatomy (Birkholz and Kröger 2006). The position and orientation of the rigid structures, as well as the shape of the deformable structures, are defined by a set of 24 vocal tract parameters. Fig. 3 illustrates the influence of the most important parameters on the shape of the vocal tract. The shape of the lips depends on the parameters *LP* and *LH* that specify the degree of protrusion and the vertical distance between the upper and lower lip. The actual geometry of the lips is calculated on the basis of the "laws for lips" by Abry and Boë (1986). The parameter *VA* determines the position and shape of the velum that can vary between a maximally raised velum and a maximally lowered velum. The actual vertex positions are calculated by interpolation between according key shapes. The vertical position of the larynx is defined by the parameter *HY*, and the horizontal position of the hyoid by *HX*. The upper part of the larynx widens or narrows according to the hyoid position, and the length of the pharynx wall is stretched or shortened according to *HY*. The mandible can be translated in the midsagittal plane by the parameters *JX* and *JY*, and can be rotated by *JA*.

Most parameters affect the shape of the tongue. In the midsagittal plane, the tongue contour is modeled by two circular arcs and two second-order Bézier curves. One of the circles represents the tongue body, and the other one the tongue tip. Their center positions are defined in Cartesian coordinates by (*TCX, TCY*) and (*TTX, TTY*). The

spline curves have two parameters each, (*TRX, TRY*) and (*TBX, TBY*), specifying the positions of their median points. In addition to the parameters defining the midsagittal tongue contour, four parameters specify the relative height of the tongue sides at four equidistant points along the contour.

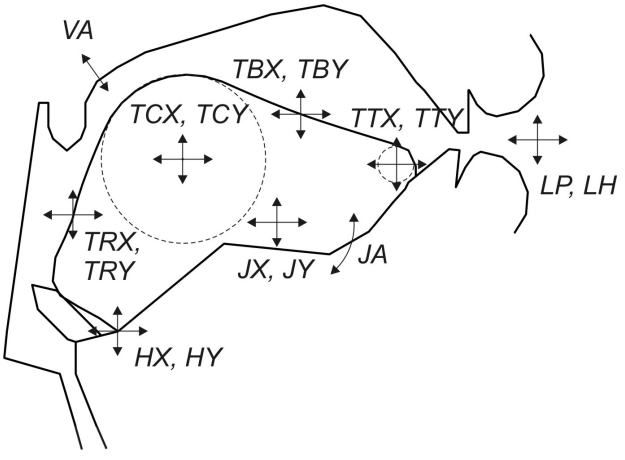

Fig. 3. Midsagittal section of the vocal tract model. The abbreviations (*TCX*, *TCY*, etc.) denote the vocal tract parameters and the arrows indicate their respective areas of influence on the model geometry.

In order to determine the vocal tract parameters for the realization of the speech sounds of Standard German, the parameters were manually adjusted in such a way that the model derived midsagittal vocal tract outlines closely matched tracings of the corresponding outlines in the MR images of the same subject as above. Fig. 4 (a), (b) and (c) illustrate the determination of MRI tracings by means of the vowel [y:]. Fig. 4 (a) shows the original midsagittal image of the vowel, and (b) shows the edges in the picture that were determined automatically by means of a Sobel operator. The vocal tract outline in (b) was then fitted with Catmull-Rom Splines that are illustrated in (c). Finally, the tracings were rotated and translated into the reference frame of the vocal tract model, and differences in head-neck angles between the subject and the vocal tract model were compensated for by warping the MRI derived outlines (Birkholz and Kröger 2006).

Fig. 5 (a) illustrates the results of the parameter tuning for the vowel [i:]. The dotted and the solid lines represent the MRI outline and the model derived outline, respectively. The dashed line represents the tongue side. Because of the flexibility of the vocal tract parameters, similar good matches could be achieved for all speech sounds in the MRI corpus. Fig. 5 (c) and (d) illustrate the determination of the vocal tract center line and the area function for a given vocal tract geometry. The centerline is calculated similarly to the method by Mermelstein (1973) and additionally smoothed. At 129 equidistant positions along the centerline, the vocal tract surfaces

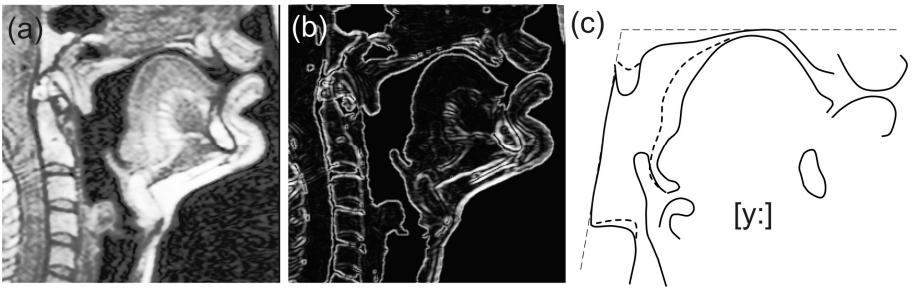

Fig. 4. (a) Midsagittal magnetic resonance image (MRI) for the phone [y:]. (b) Edges detected by a Sobel operator for (a). (c) Tracing results for (b). The thick dashed lines indicate the left side of the tongue.

are intersected with planes perpendicular to the centerline. The areas of the resulting cross-sections are calculated numerically and constitute a high-resolution area function. This area function is low-pass filtered and mapped on the cross-sectional areas and perimeters of 40 abutting elliptical tube sections that comprise the tube model of the vocal tract.

To test the *acoustic* match between the vowels spoken by the MRI subject and the synthesized vowels with the tuned parameters, the first three formants of all vowels in the corpus were compared. The formants of the subject's vowels were determined by a LPC analysis. The formants of the synthetic vowels were extracted from a frequency domain simulation of the vocal system on the basis of the model derived area functions (Birkholz and Jackèl 2004). The error between the natural and synthetic formant frequencies averaged over all vowels was about 12%. This error must be mainly attributed to the resolution-limited accuracy of the MRI tracings as well as to the imperfect matching of the outlines. It is well known that in certain regions of the vocal tract, the formant frequencies are quite sensitive to small variations of articulatory parameters (Stevens 1989). Therefore, the acoustic differences could be caused by only small articulatory deviations due to the above sources of errors. To test how far small corrective variations of the vocal tract parameters can improve the acoustic match, we implemented an algorithm searching the parameter space to minimize the formant errors. With only little articulatory changes of the vocal tract parameters in the sensitive regions, the average formant error reduced to 3.4% (Birkholz and Kröger 2006).

For the voiced excitation of the synthesizer, we implemented a parametric model of the glottal geometry based on Titze (1984). A schematic representation of the model is shown in Fig. 6. The vocal fold parameters are the degree of abduction at the posterior end of the folds at the lower and upper edge, the fundamental frequency F_0, the phase difference between the upper and lower edge, and the subglottal pressure. Based on these parameters, the model generates the time-varying cross-sectional areas at the glottal inlet and outlet opening that are in turn mapped on two glottal tube sections (an upper and a lower one). We extended Titze's original model to account for a smooth diminishment of the oscillation amplitude with increasing abduction and for a parametrization of glottal leakage similar to Cranen and Schroeter (1995).

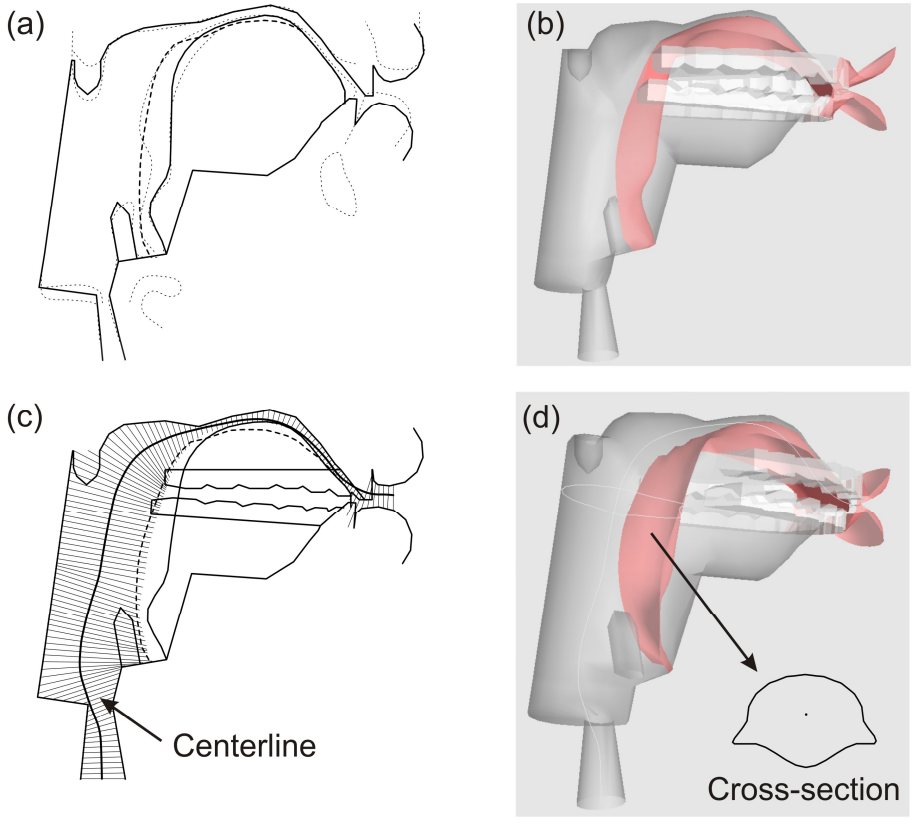

Fig. 5. (a) The vocal tract parameters were adjusted in such a way that the model outline (solid lines) closely matches the MRI tracing for the vowel [i:] (dotted lines). (b) 3D rendering of the vocal tract model for the adapted vowel [i:]. (c) Midsagittal section of the vocal tract model with the centerline. The straight lines normal to the centerline indicate the positions and orientations of the cutting planes. (d) Centerline and cross-section in the pharyngeal region within the 3D geometry of the vocal tract.

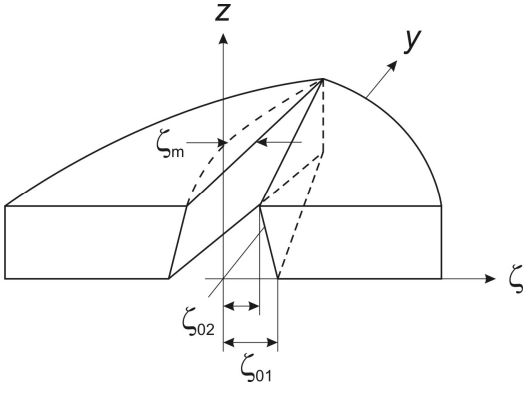

Fig. 6. Model of the vocal folds/glottis based on Titze (1984)

2.2 Acoustic Simulation

To generate the actual acoustic signal, the geometric models of the vocal tract and the vocal folds are transformed into a composite discrete area function and combined with area functions for the trachea, and the nasal tract with paranasal sinuses (according to Dang and Honda 1994). These area functions represent the branched tube model of the whole vocal system as illustrated in Fig. 7 (top). The areas of the

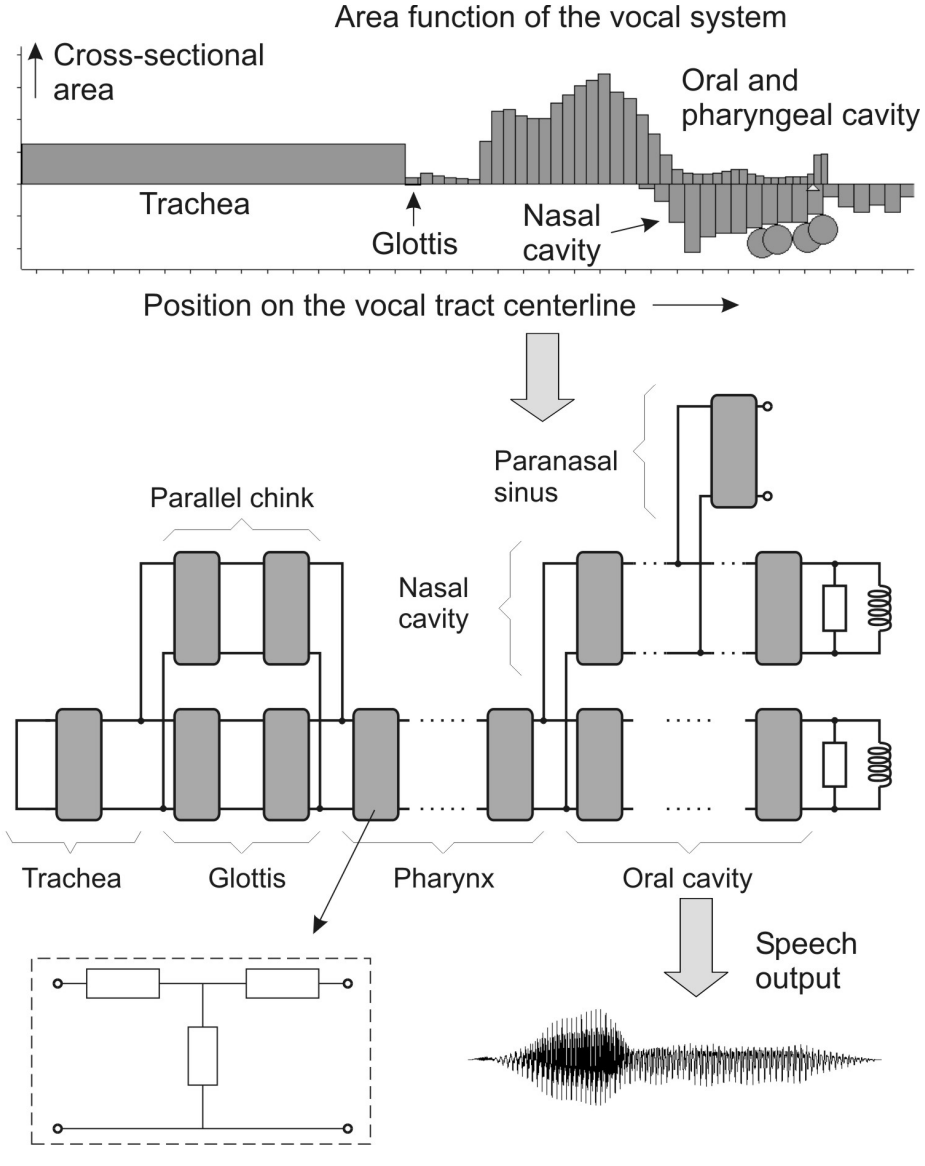

Fig. 7. Generation of the acoustic speech signal (see text)

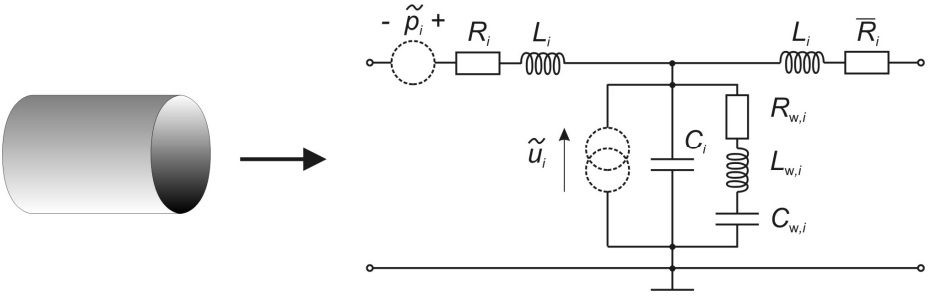

Fig. 8. Correspondence between a short tube section and a two-port network section of an electrical transmission line with lumped elements

nasal cavity are flipped upside down and the paranasal sinuses are drawn as circles. This tube model is then transformed into an equivalent electrical transmission line network as in Fig. 7 (middle). In this network, each tube section is represented by a two-port section as depicted in Fig. 8. On the basis of electro-acoustic analogies, the geometric parameters of the tube sections are transformed into the electrical network components of the two-port sections (Birkholz and Jackèl 2004, Birkholz 2005, Birkholz et al. 2007a). In this network, electrical current corresponds to volume velocity and voltage corresponds to sound pressure. The time-varying distribution of volume velocity and pressure in the tube model is simulated in the time domain based on finite difference equations for the corresponding variables in the network at a sampling rate of 44.1 kHz. Specific algorithms were implemented for the simulation of losses due to friction, sound radiation, and wall vibration, as well as for the generation of noise due to turbulence. The radiated speech signal is approximated as the time derivative of the volume velocity through the mouth opening and the nostrils. At each time instant of the simulation, a new set of parameters for the vocal tract and the vocal folds can be specified, leading to an updated tube model geometry which leads to changes in the acoustic properties of the vocal system. The implemented acoustic simulation is capable of generating speech sounds of all phonetic categories (e.g. vowels, fricatives, plosives, nasals, …) on the basis of appropriate area functions.

In summary, the presented synthesizer does not only generate a high quality acoustic output, but also the corresponding sequence of 3D vocal tract shapes and area functions. The geometrical output can, for example, be used for the animation of the lip and chin region of a face model of a virtual speaker (avatar). Furthermore both outputs – geometrical information of articulator positions (e.g. palatal contact patterns) and the acoustic speech signal – are essential for learning speech production (motor execution for speech articulation) and may help refining speech perception (place of articulation, see motor theory of speech perception).

3 The Gesture-Based Control Concept

The term *gesture* as used in this paper denotes the concept of *target-directed speech movements realizing speech sounds* (e.g., a lip closing gesture in temporal

coordination with a glottal opening-closing gesture for the realization of a voiceless labial stop /p/). From the viewpoint of motor control, gestures are target-directed actions, i.e., prelearned high-level motor programs accomplishing a specific task like lip closure or glottal opening without specifying the concrete muscular activity for each motor unit on the primary motor level (Saltzman and Munhall 1989, Ito et al. 2004). This concept of the *gesture* or *motor action* is widely accepted for movement control of biological systems in general - not only for speech movement control, but also for nonverbal gestures like eye-brow movements as well as for head, arm, and other (speech-accompanying) body movements (Sober und Sabes 2003, Todorov 2004, Paine und Tani 2004, Fadiga und Craihgero 2004).

The gesture-based control concept introduced here is related to the control concept introduced by Saltzman and Munhall (1989) which was extended to a linguistic theory by Browman and Goldstein (1989, 1990, and 1992). A controversial discussion of this concept was stimulated by these papers, mainly due to problems concerning the quantitative implementation of this approach (e.g., Kröger et al. 1995) and due to the fact that perception as a control instance for production is not considered in this approach (Kohler 1992). Especially the control of gestural target positions is actually mainly done in perceptual domains (Perkell et al. 1997, Guenther et al. 1998, Guenther and Perkell 2004).

On the other hand, there are a lot of well-accepted facts privileging a gestural concept above simpler segmental target concepts for describing speech movements and for establishing speech motor control models. In the case of vowels it is well known that their targets need not to be reached completely for producing or perceiving a vowel phoneme. In the case of casual speech or in the case of high speaking rate, vowel realizations often exhibit target undershoot without affecting the perceptual identification and discrimination of the phoneme produced (Lindblom 1963, Strange 1989). That indicates the importance of formant trajectories compared to steady state formant patterns and thus the importance of target-directed articulatory movements instead of pure articulatory target locations. This supports the hypothesis that vocalic targets are perceived rather from the acoustic results of articulatory movements towards an auditory target than from the real achievement of the (articulatory or acoustic) target positions. This underlines that articulatory movement transitions are more important for speech perception than reaching the target positions and thus supports the concept of the speech gesture.

In addition, the gesture-based approach introduced in this paper does not longer exhibit shortcomings of earlier gestural concepts: (i) The quantitative model for describing goal-directed movements is changed; (ii) gesture execution is learned using external auditory and sensory (auditory and somatosensory) feedback information; (iii) learning concepts are used to implement the complete repertoire of speech gestures for a specific language.

The basic assumptions of the gestural concept introduced in this paper are:

(i) Each speech sound (or phoneme realization) consists of one or more gestures distributed over gestural tiers (Tab. 1 and Fig. 9). Vocalic sounds are composed of a tract-shaping gesture determining the vowel quality (high-low, front-back, rounded-unrounded), a glottal closing gesture responsible for voicing (clgl, see Tab. 1), and a velopharyngeal closing gesture (clvp) ensuring that the vowel is not nasalized. The realization of a consonant phoneme is composed of a consonantal near closing or

(full)closing gesture determining manner and place of articulation (see Tab. 1), a glottal opening or closing gesture determining whether the consonant is voiced or unvoiced, and a velopharyngeal opening (opvp) or tight closing gesture (tcvp) determining whether the sound is nasal or oral.

(ii) A temporal interval of activation is defined for each gesture (Fig. 9, shaded boxes). This temporal interval comprises a transition portion and (optionally) a target portion if the target is reached (both portions are not separated strictly in Fig. 9). Target portions occur in many types of gestures when articulatory saturation effects occur (e.g. contact of two articulators or contact of an articulator with the vocal tract walls, Kröger 1993 and Kröger 1998). Saturation effects lead to constant effective vocal tract constrictions (or glottal or velopharyngeal constriction) over the target portion time interval while the vocal tract parameter values continue to change towards the target (see non-dashed and dashed line in Fig. 10). For example in the case of contact between the tongue tip and the hard palate during a consonantal apical closing gesture, a time interval for complete closure occurs. Within this time interval the contact area between the articulator and the vocal tract wall changes but the complete closure remains.

Table 1. List of vocalic, consonantal, velopharyngeal, and glottal speech gestures (not complete). Default gestures are marked by an asterisk. Keep in mind that gestures should not be confused with phonemes (e.g., the phoneme /b/ is realized by a {b}-gesutre in combination with a {pho}-gesture; the phoneme /p/ is realized by a {b}- in combination with a {-pho}-gesture; the phoneme /m/ is realized by a combination of a {b}-, a {nas}-, and a {pho}-gesture). Moreover the gestural symbols represent distinctive features or bundles of distinctive features realized by this gesture. The symbol {-nas} means non-nasalized sonorant. The symbol {obs} means obstruent, which is by definition non-nazalised sound and which always needs a velopharyngeal tight closure in order to avoid pressure loss in the mouth cavity.

Abbreviation	*Symbol*	*Name of Gesture*
iivt	{i:}	(vocalic) vocal tract ii-shaping
uuvt	{u:}	(vocalic) vocal tract uu-shaping
aavt	{a:}	(vocalic) vocal tract aa-shaping
swvt *	{schwa} *	(vocalic) vocal tract schwa-shaping *
….	….	….
clla	{b}	(consonantal) labial closing
clap	{d}	(consonantal) apical closing
cldo	{g}	(consonantal) dorsal closing
ncld	{v}	(consonantal) labio-dental near closing
ncal	{z}	(consonantal) alveolar near closing
ncpo	{Z}	(consonantal) postalveolar near closing
….	….	….
clvp *	{-nas} *	velopharyngeal closing *
tcvp	{obs}	velopharyngeal tight closing
opvp	{nas}	velopharyngeal opening
….	….	….
clgl *	{pho} *	glottal closing *
tcgl	{?}	glottal tight closing
opgl	{-pho}	glottal opening
….	….	….

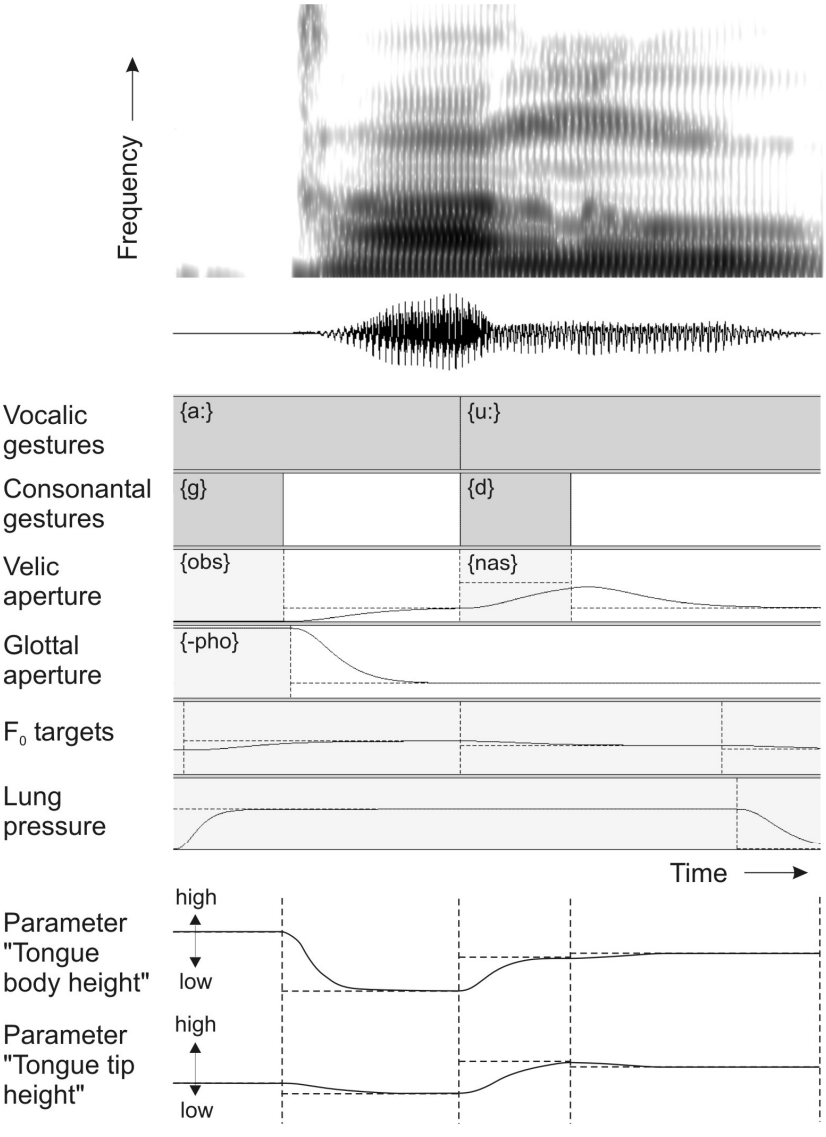

Fig. 9. Gestural score for the utterance /ka:nu:/: Six tiers are given: vocalic tier, consonantal tier, velic tier, glottal tier, fundamental frequency tier, and lung pressure tier. Gestural activation is marked by the shaded boxes. In the case of velic, glottal, and lung pressure gestures only one parameter is controlled (e.g. *VA* for a velic gestures). In the case of vocalic and consonantal gestures many vocal tract paratemers are controlled. As an example, the spatio-temporal trajectories for the vocal tract parameters tongue body height *TBH* and tongue tip height *TTH* are displayed. In the case of both parameters the targets for the {g}-, {a:}-, {d}-, and {u:}-gesture are displayed (dashed lines). If no activation is marked on the vocalic, velic, or glottal tier, the appropriate default gesture becomes activated. If no activation is marked on the consonantal tier, no consonantal gesture is active. The vocal tract is solely determined by vocalic gestures in this case.

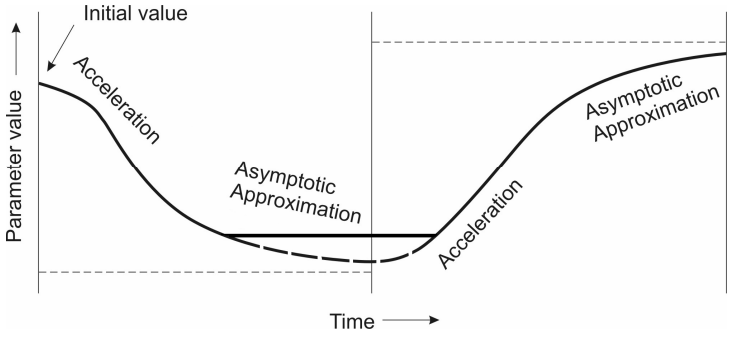

Fig. 10. Spatiotemporal control parameter trajectory for an articulator movement in the case of two successive gestures. Durations and target locations for the gestures are indicated by thin dashed lines. The bold horizontal line indicates a saturation effect (i.e., a constant degree of constriction) while the control parameter further changes (bold dashed line). The distance between the dashed bold line and the solid bold line indicates the degree of articulator to articulator or articulator to vocal tract wall contact.

(iii) Gestural time functions are quantitatively described by a third order cascaded system (Birkholz 2007c and cf. Ogata and Sonoda 2000) which is capable of approximating natural articulatory data better than the critically damped second order system used by Browman and Goldstein (1990b) and Smith et al. (1993) (see also Kröger et al. 1995). Within the quantitative approach by Birkholz (submitted) gestural rapidity (i.e., the time period of the gestural transition portion or the time period needed for approximating the gestural target) is defined controlling the articulatory velocity for each gesture. Target values and gestural rapidity can vary for each gesture, for example with respect to different prosodic conditions of the corresponding syllable.

(iv) In many cases more than one articulator is involved in the gestural movement realization (e.g. upper lips, lower lips, and mandible in the case of a labial closing gesture). A concept of articulatory dominance is used to determine the contribution of each articulator for the realization of a gestural movement (Birkholz 2007c, Birkholz et al. 2006). Also in the case of temporal overlap of gestures on the same or on different tiers, the resulting gestural movement trajectories for each articulator are calculated using the dominance concept.

4 Learning Gesture-Based Control Using Sensory Feedback and External Speech Signals

Over the last decades it became more and more apparent that purely rule-driven formant synthesis systems are not suitable for producing natural sounding speech (i.e., high-quality speech signals) in comparison to data-driven concatenative speech synthesis like the unit-selection approach. Thus, a rule based gestural control module for articulatory speech synthesis does not lead automatically to high quality synthetic speech. To overcome this problem, one idea is to combine unit selection approaches with articulatory speech synthesis (Birkholz et al. 2007b), e.g., to generate crucial

control information like segmental durations using unit selection modules. The idea introduced here is different. If we want to model speech production and speech perception as close as possible to real biological processes, it should be tried to extract control rules from learning procedures as toddlers do it during speech acquisition. That means to extract control information (i) from audible feedback signals during self-productions (babbling) and (ii) from external auditory and visual speech signals by imitation.

An outline of learning steps that we implemented to establish control rules for the gesture-based articulatory speech synthesizer is given below (not complete):

(i) In order to cope with the many-to-one articulatory-to-acoustics inversion problem in vowel production, constraints must be introduced regarding the possible articulator configurations for vowel targets (e.g., constraints for jaw position in isolated vowel production). This subset of articulatory configurations must still cover the whole vowel space, i.e., the dimensions low-high, front-back, and rounded-unrounded. It can be learned or trained during babbling (Kröger et al. 2006b).

(ii) In order to consider coarticulatory effects of neighboring consonants on the articulation of vowels (e.g., to allow higher jaw positions for a vowel produced in the context of alveolar fricatives compared to isolated production) rules for compensatory articulation have to be learned during babbling such that the production model is capable of producing comparable vocal tract shapes (or area functions) using different (compensatory) articulator positions (e.g., a decrease in jaw position has to be compensated for by an increase in tongue elevation and lower lips elevation, and vice versa, see Kröger et al. 2006c).

(iii) The production of consonantal closing gestures and the perception of the place of articulation is also learned during babbling in our approach using a V_1CV_2 training set where V_1 and V_2 cover the whole vowel space and C covers lip, tongue tip, and tongue body closures (Kröger et al. 2006a and 2006b). This learning step leads to a database combining perceptual knowledge and articulatory knowledge – i.e. which formant transition in which vocalic context is produced for each place of articulation (labial, dental, alveolar, postalveolar, palatal, velar).

(iv) The production of other types of consonants like fricatives and nasals is also learned during babbling in our approach. For learning optimal friction noise production the degree of constriction is varied at different places of articulation. For learning the difference between nasal and obstruent sound production, different maneuvers of the velum (i.e., wide opening vs. tight closure) are trained in parallel to the oral closing gesture.

(v) During babbling, the model mainly acquired general, language independent phonetic knowledge, like the relation between formants and articulatory positions for vocalic sounds, the relation between formant transitions and consonantal closing gestures in different vocalic contexts, and so on. After this, the model is ready to imitate external productions, i.e., to learn from acoustic and optical data produced by other speakers. This means to learn language specific speech production and in consequence language specific perception, e.g., the perceptual magnet effect (Kröger et al. 2007).

(vi) The next steps of imitation training are (a) to strengthen the articulation of language specific places and manners of articulation, (b) to learn the language specific gestural coordination for consonant clusters and for different syllable structures

(simple to complex), and (c) to learn language specific prosodic items like intonation patterns. These training steps have not been modeled so far.

5 Discussion

An articulatory speech synthesizer comprising a three-dimensional vocal tract model and a gesture-based concept for the control of articulatory movements has been described in this paper. The concept of speech gestures for quantitative control of speech articulation introduced here is discussed carefully with respect to other gestural approaches. The gestural model introduced here comprises a quantitative specification for sound targets and for the generation of the spatiotemporal gestural movement trajectories. It has been emphasized that a quantitative concept for describing goal-directed actions is very important because these trajectories define the perceptually important formant trajectories. A modular learning strategy based on speech perception is outlined for the built-up of gestural control rules. Sensory feedback information for defined articulatory items produced by the model itself and auditory and visual information of speech items produced by external speakers are used successively for the generation of the control rules. The complete model (control module and synthesizer) is capable of producing high-quality synthetic speech signals and gives a scheme for speech production and speech perception which is closely related to the human biological speech production and perception mechanisms. Potential applications for this model are (i) high-quality speech synthesis, (ii) establishing models for human speech production and perception, capable of helping to understand normal speech production, normal speech perception, and in addition, the underlying mechanisms of motor speech disorders (Kent 2000).

Acknowledgments. This work was supported in part by the German Research Council DFG grants JA 1476/1-1 and Kr 1439/10-1.

References

Abry, C., Boë, L.J.: Laws for lips. Speech Communication 5, 97–104 (1986)

Birkholz, P.: 3D-Artikulatorische Sprachsynthese. Unpublished PhD thesis. University Rostock (2005)

Birkholz, P.: Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In: Proceedings of the Interspeech 2007 - Eurospeech. Antwerp, Belgium (2007c)

Birkholz, P., Jackèl, D.: Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In: Proceedings of Interspeech 2004-ICSLP. Jeju, Korea, pp. 1125–1128 (2004)

Birkholz, P., Kröger, B.J.: Vocal tract model adaptation using magnetic resonance imaging. In: Proceedings of the 7th International Seminar on Speech Production, pp. 493-500. Belo Horizonte, Brazil (2006)

Birkholz, P., Jackèl, D., Kröger, B.J.: Construction and control of a three-dimensional vocal tract model. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), Toulouse, France, pp. 873-876 (2006)

Birkholz, P., Jackèl, D., Kröger, B.J.: Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15, 1218–1225 (2007a)

Birkholz, P., Steiner, I., Breuer, S.: Control concepts for articulatory speech synthesis. In: Proceedings of the 6th ISCA Speech Synthesis Research Workshop. Universität Bonn (2007b)

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

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

Browman, C.P., Goldstein, L.: Gestural specification using dynamically-defined articulatory structures. Journal of Phonetics 18, 299–320 (1990b)

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

Cranen, B., Schroeter, J.: Modeling a leaky glottis. Journal of Phonetics 23, 165–177 (1995)

Dang, J., Honda, K.: Morphological and acoustical analysis of the nasal and the paranasal cavities. Journal of the Acoustical Society of America 96, 2088–2100 (1994)

Fadiga, L., Crahighero, L.: Electrophysiology of action representation. Journal of clinical Neurophysiology 21, 157–169 (2004)

Flanagan, J.L.: Speech Analysis, Synthesis and Perception. Springer, Berlin (1965)

Guenther, F.H., Perkell, J.S.: A neural model of speech production and its application to studies of the role of auditory feedback in speech. In: Maassen, B., Kent, R., Peters, H., van Lieshout, P., Hulstijn, W. (eds.) Speech motor control in normal and disordered speech, pp. 29–49. Oxford University Press, Oxford (2004)

Guenther, F.H., Hampson, M., Johnson, D.: A theoretical investigation of reference frames for the planning of speech movements. Psychological Review 105, 611–633 (1998)

Guenther, F.H.: Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39, 350–365 (2006)

Guenther, F.H., Ghosh, S.S., Tourville, J.A.: Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language 96, 280–301 (2006)

Ito, T., Gomi, H., Honda, M.: Dynamic simulation of speech cooperative articulation by muscle linkages. Biological Cybernetics 91, 275–282 (2004)

Kent, R.D.: Research on speech motor control and its disorders: A review and prospective. Journal of Communication disorders 33, 391–428 (2000)

Kohler, K.J.: Gestural reorganization in connected speech: A functional viewpoint on 'articulatory phonology'. Phonetica 49, 205–211 (1992)

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

Kröger, B.J.: Ein phonetisches Modell der Sprachproduktion. Niemeyer Verlag, Tübingen (1998)

Kröger, B.J., Birkholz, P., Kannampuzha, J., Neuschaefer-Rube, C.: Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer. In: Proceedings of the 9th International Conference on Spoken Language Processing, Interspeech 2006, ICSLP, pp. 565–568 (2006a)

Kröger, B.J., Birkholz, P., Kannampuzha, J., Neuschaefer-Rube, C.: Learning to associate speech-like sensory and motor states during babbling. In: Proceedings of the 7th International Seminar on Speech Production. Belo Horizonte, Brazil, pp. 67–74 (2006b)

Kröger, B.J., Birkholz, P., Kannampuzha, J., Neuschaefer-Rube, C.: Spatial-to-joint coordinate mapping in a neural model of speech production. In: DAGA-Proceedings of the Annual Meeting of the German Acoustical Society. Braunschweig, Germany, pp. 561–562 (2006c)

Kröger, B.J., Birkholz, P., Kannampuzha, J., Neuschaefer-Rube, C.: Modeling the perceptual magnet effect and categorical perception using self-organizing neural networks. In: Proceedings of the International Congress of Phonetic Sciences. Saarbrücken, Germany (2007)

Kröger, B.J., Schröder, G., Opgen-Rhein, C.: A gesture-based dynamic model describing articulatory movement data. Journal of the Acoustical Society of America 98, 1878–1889 (1995)

Lindblom, B.: Spectrographic study of vowel reduction. Journal of the Acoustical Society of America 35, 1773–1781 (1963)

Mermelstein, P.: Articulatory model for the study of speech production. Journal of the Acoustical Society of America 53, 1070–1082 (1973)

Ogata, K., Sonoda, Y.: Evaluation of articulatory dynamics and timing based on cascaded first-order systems. In: Proceedings of the 5th Seminar on Speech Production, Kloster Seeon, Germany, pp. 321–324 (2000)

Paine, R.W., Tani, J.: Motor primitive and sequence self-organization in a hierarchical recurrent neural network. Neural Networks 17, 1291–1309 (2004)

Perkell, J.S., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., Guiod, P.: Speech motor control: Acoustic goals, saturaltion effects, auditory feedback and internal models. Speech communication 22, 227–250 (1997)

Saltzman, E.L., Munhall, K.G.: A dynamic approach to gestural patterning in speech production. Ecological Psychology 1, 333–382 (1989)

Smith, C.L., Browman, C.P., Kay, B., McGowan, R.S.: Extracting dynamic parameters from speech movement data. Journal of the Acoustical Society of America 93, 1580–1588 (1993)

Sober, S.J., Sabes, P.N.: Multisensory integration during motor planning. The Journal of Neuroscience 23, 6982–6992 (2003)

Stevens, K.N.: On the quantal nature of speech. Journal of Phonetics 17, 3–45 (1989)

Strange, W.: Dynamic specification of coarticulated vowels spoken in sentence context. Journal of the Acoustical Society of America 85, 2135–2153 (1989)

Titze, I.R.: Parameterization of the glottal area, glottal flow, and vocal fold contact area. Journal of the Acoustical Society of America 75, 570–580 (1984)

Todorov, E.: Optimality principles in sensorimotro control. Nature Neuroscience 7, 907–915 (2004)