Control of an Articulatory Speech Synthesizer based on Dynamic Approximation of Spatial Articulatory Targets

Peter Birkholz

Institute for Computer Science, University of Rostock
Albert-Einstein-Str. 21, 18059 Rostock, Germany
piet@informatik.uni-rostock.de

Abstract

We present a novel approach to the generation of speech movements for an articulatory speech synthesizer. The movements of the articulators are modeled by dynamical third order linear systems that respond to sequences of simple motor commands. The motor commands are derived automatically from a high level schedule for the input phonemes. The proposed model considers velocity differences of the articulators and accounts for coarticulation between vowels and consonants. Preliminary tests of the model in the framework of an articulatory speech synthesizer indicate its potential to produce realistic speech movements and thereby to contribute to a higher quality of the synthesized speech.

Index Terms: Speech synthesis, vocal tract model, articulatory control, target approximation

1. Introduction

An important part of each articulatory speech synthesizer is a method for the generation of the artificial speech movements. A realistic simulation of the spatial and temporal properties of speech movements is crucial for high quality articulatory speech synthesis, because the human ear is very sensitive to the dynamic aspects of speech. In this paper, we present a novel model for the quantitative control of supraglottal articulatory movements on the basis of a high level parameterized "phoneme schedule" of the intended utterance. The presented model was implemented as part of a comprehensive articulatory speech synthesizer based on a 3D model of the vocal tract and a time-domain simulation for the aeroacoustic generation of speech sounds [1, 2, 3].

Currently, there are only few models for the control of articulatory synthesizers (as there are only few articulatory synthesizers at all). They either model directly the surface countour of speech movements by interpolation (e.g., Mermelstein [4] and the kinematic model by Kröger [5]), or they try to simulate the underlying mechanisms of speech production. Examples for the latter approach are the task dynamic model by Saltzman and Munhall [6] and the gesture-based dynamic model by Browman and Goldstein [7].

The proposed control model also belongs to the second category, but differs in most respects from the aforementioned approaches, in particular in the definition and execution of motor commands. A flow chart of the model is illustrated in Fig. 1. The input to the system is a parameterized schedule of the phonemes to be articulated. This schedule defines temporal intervals for the phonemes, the temporal overlap between vowel and consonants, and the "articulatory effort" for their realization. In this respect, the schedule is similar to gestural scores as

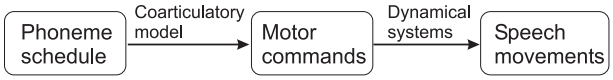

Figure 1: Flow chart of the proposed control model.

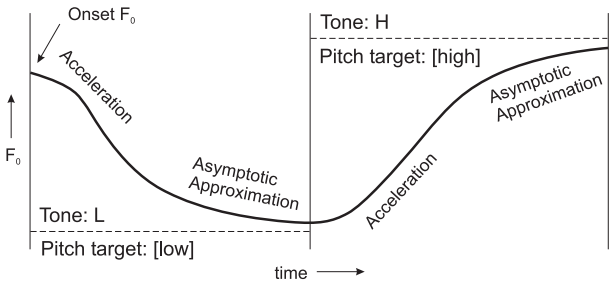

Figure 2: A schematic sketch of the Target Approximation model (by Xu and Wang [8]). The vertical lines represent syllable boundaries and the dashed lines represent underlying pitch targets. The thick curve represents the F_0 contour that results from articulatory implementation of the pitch targets.

defined by Browman and Goldstein [7]. The schedule is transformed into motor commands by means of prototypical articulatory phoneme targets and a dominance model for the consideration of coarticulatory effects. For each control parameter of the vocal tract model, one sequence of motor commands is generated. A motor command is implemented as a target position for its associated vocal tract parameter within a defined temporal interval. The actual motion of the articulators, i.e., the temporal change of the vocal tract parameters, result from the dynamic successive approximation of the assigned target positions.

The definition of the motor commands and their execution by means of dynamical systems was inspired by the target approximation (TA) model for F_0 production by Xu and Wang [8]. In this model, surface F_0 contours result from asymptotic approximations of underlying pitch targets as illustrated in Fig. 2. Recent experiments by Xu and Liu [9] suggested that also the movement of other speech articulators could be explained as a process of sequential target approximation. The proposed model contains a first quantitative implementation of this idea as described in Section 2. Section 3 presents the method for the specification of the motor commands by means of a phoneme schedule. A brief discussion follows in Section 4.

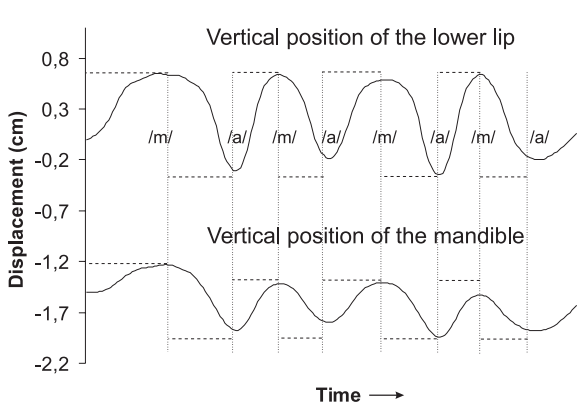

Figure 3: Real movement traces of the vertical position of the lower lip and the mandible during the utterance /mamamama/. The dashed lines represent the assumed targets for these variables in the spirit of the proposed model. The dotted lines are the boundaries between the hypothesized motor commands.

2. Motor commands and their dynamic execution

In the context of the proposed model, a (simple) motor command is defined as a target value for an articulatory position variable within a defined time interval. Let $y(t)$ denote such a position variable of an arbitrary movable structure of the vocal tract, e.g., the vertical position of the lower lip or the tongue tip in a fixed reference frame. To get a solution for $y(t)$, we need an equation that describes the motion of the associated articulator as response to a given a sequence of motor commands. As an example, Fig. 3 shows the articulatory traces of two position variables recorded by an electromagnetic articulograph (EMA) for the utterance /mamamama/ (stress on the third syllable). The upper curve shows the vertical position of the lower lip, and the lower curve shows the vertical position of the mandible. The equation of motion controlling the respective articulator positions in our model should be able to replicate such measurements as close as possible on the basis of a given sequence of motor commands.

The dotted lines and the dashed lines in Fig. 3 illustrate our idea of the temporal boundaries and the target positions of the motor commands (inspired by [9]). Note that for each variable the target positions are assumed to be equal for all instances of /m/ and /a/ in this example (except for the jaw variable for the initial /m/). However, the variables do not always reach their targets within the interval of the motor commands. To what degree an articulatory variable reaches its target depends on the dynamics of the corresponding articulator and the time available for the execution of the command. Note furthermore that the turning points of the movement curves do not necessarily coincide with the command boundaries. It is much more likely that the turning points lie a bit *behind* the boundaries due to the inertia of the articulators.

Concerning the equation of motion for the articulatory variables, a natural choice seems to be that of a second order linear system, when we presume an analogy with a damped mass-spring system:

$$m\ddot{y} + r\dot{y} + k(y - y_0) = 0. \tag{1}$$

Here, m, r, k, and y_0 represent the mass of the movable structure, its internal friction, the spring constant, and the underlying target position, respectively. In order to avoid a target overshoot with such a system, which would violate the assumption of *unidirectional* target approximation, it must be critically damped or overdamped. A critically damped second order dynamic system is for example used by Saltzman and Munhall [6]. However, Kröger *et al.* [10] observed that – excited by step signals – such systems can not fit real articulatory traces with high accuracy. This becomes immediately plausible when you compare the step response of a critically damped second order system (Fig. 4, leftmost curve) with the movement traces in Fig. 3. While the real movements from one target to the next look similar to half-cycles of an undamped oscillation, the "S" like movement curve of the *critically damped* second order system has a rather rapid acceleration phase, but a much smoother deceleration phase.

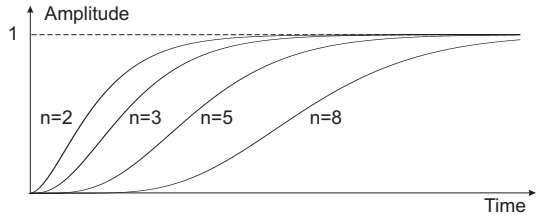

Figure 4: Step responses for cascaded first order systems of 2nd, 3rd, 5th, and 8th order.

Ogata and Sonoda [11] propose to use cascaded first order systems to describe articulatory activities and demonstrated their ability to fit natural movement traces with very high accuracy. The transfer function of such a system with n cascaded equal components is

$$H(s) = 1/(1 + \tau s)^n, \tag{2}$$

where s is the complex frequency and τ is the time constant. Figure 4 shows the step responses of cascaded first order systems for $n = 2$, 3, 5, and 8. For $n = 2$, the system corresponds to the critically damped second order system. With increasing order, the step response approaches the shape of the Gaussian cumulative distribution function.

In this study, we use a third order cascaded system to simulate articulatory movements. Therefore, the equation of motion for an articulatory variable within the temporal interval of a motor command reads

$$\tau_i^3 \dddot{y}_i + 3\tau_i^2 \ddot{y}_i + 3\tau_i \dot{y}_i + y_i = b_i, \tag{3}$$

where t is the time since the beginning of the interval, and b_i is the target position of the motor command. The only adjustable parameter next to b_i is the time constant τ_i, which can be interpreted as the inverse of the "articulatory effort". The smaller τ_i, the faster is the approach towards the underlying target. The solution of Eq. (3) for $y_i(t)$ is

$$y_i(t) = (c_{1,i} + c_{2,i}t + c_{3,i}t^2)e^{-t/\tau_i} + b_i. \tag{4}$$

The coefficients $c_{1,i}$, $c_{2,i}$, and $c_{3,i}$ can be determined from the continuity constraint for the position $y(t)$, the velocity $\dot{y}(t)$ and the acceleration $\ddot{y}(t)$ at the boundary between the motor commands. Beginning with the first motor command in a sequence, the curve sections $y_i(t)$ can be successively calculated for all subsequent commands.

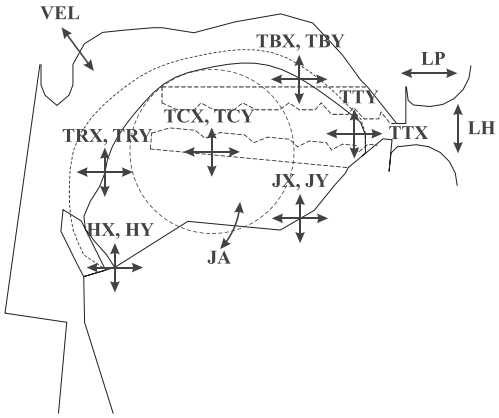

Figure 5: Schematic overview of the parameters of the vocal tract model and the articulatory structures that they control.

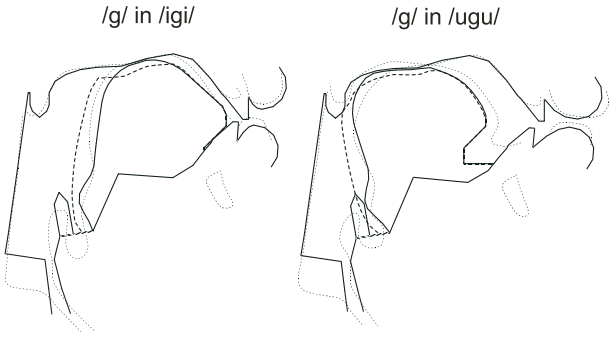

Figure 6: Different realizations of the consonant /g/ in the context of the vowels /i/ (left) and /u/ (right) due to coarticulation. The dotted contours are tracings of dynamic magnetic resonance images, and the solid lines show the corresponding model generated contours.

3. High-level prediction of motor commands

In this section, we briefly introduce the vocal tract model and specify, how the motor command sequences for all of its parameters can be generated on the basis of a parameterized phoneme schedule.

3.1. Vocal tract model

The vocal tract model of our synthesizer is a three-dimensional wire frame representation of the surfaces of the articulators and the vocal tract walls of a male speaker [1, 2]. The shape and position of all movable structures is a function of 23 adjustable parameters. The most important parameters and their respective areas of influence are sketched in the midsagittal section of the model in Fig. 5. Some of the parameters come in pairs and define the position of certain structures directly in cartesian coordinates in the reference frame of the hard palate. For example, (*TCX,TCY*) defines the position of the tongue body (represented by a circle), (*TTX,TTY*) defines the position of the tongue tip, and (*JX,JY*) the position of the mandible. Therefore, the temporal change of these parameters should closely reflect articulatory movements measured by articulographic devices, e.g., EMA.

Recently, the parameter values of the model have been adjusted for all German phonemes by means of volumetric magnetic resonance images [2]. Hence, we know the target configuration for all German vowels and consonants in terms of parameter values. The target configurations for consonants define the static vocal tract shape at the time of maximum constriction without a specific phonetic context. However, it is well known that the actual articulatory realization of consonants is strongly influenced by its vocalic context, e.g., a /g/ in /igi/ is realized differently from /g/ in /aga/. To represent these coarticulatory differences, we defined a dominance value for each parameter of each consonant. A high dominance means that the parameter is important for the formation of the consonantal constriction (e.g., the vertical tongue tip position for /d/). On the other hand, parameters with a low dominance are influenced by the vocalic context (e.g., the horizontal tongue body position for /g/). Formally, this concept is expressed by

$$x_{c|v}[i] = x_v[i] + w_c[i] \cdot (x_c[i] - x_v[i]), \qquad (5)$$

where i is the parameter index, $x_{c|v}[i]$ is the value of parameter i at the moment of the maximal closure/constriction of the consonant c in the context of the vowel v, $w_c[i]$ is the weight/dominance of parameter i, and $x_c[i]$ and $x_v[i]$ are the parameter values of the targets for the consonant and vowel. The optimal dominance values for all parameters of all consonants have been determined in a previous study [2]. It was also shown that this simple dominance model is capable of reproducing the major coarticulatory differences in the production of consonants. As an example, Fig. 6 shows the different realization of /g/ in /igi/ and /ugu/ using this dominance model. The dotted contours are MRI tracings of the midsagittal vocal tract during the production of /g/ in the corresponding contexts.

3.2. Motor command specification

According to Sec. 2, the proposed control model generates the movement of each vocal tract parameter by a sequence of simple motor commands. In this section, we present a concept to predict the motor command parameters – its time interval, target, and "articulatory effort" – on the basis of a parameterized phoneme schedule. The main assumption of that concept is that all (supraglottal) articulators start their motion towards a new target position at the same time. Hence, the boundaries between the motor commands coincide for all vocal tract parameters. Indications of this behavior are summarized in [9] and are also evident for the EMA traces in Fig. 3. An exception to this rule in our model is the parameter for the velic aperture, which must be controlled by an independent motor command sequence. Furthermore, we assume that the *relative* speed between different articulators is constant during an utterance. Mermelstein [4], for example, observed that articulatory variables involved in the formation of a consonantal constriction undergo an exponential-like transition following the release, and that the tongue body, lips, tongue tip, and jaw differ in the time constant of that transition. He found a time constant of roughly 75 ms for the tongue body and the jaw, 30 ms for the lips, and 50 ms for the tongue tip. In our model, we assume that these relative velocity differences reflect inherent properties of the articulators and are valid for any kind of articulatory movement. This means, e.g., that the opening or closing of the lips is always executed 75/30=2.5 times as fast as a simultaneous translation of the jaw. In summary, the above assumptions make all model articulators start at the same time towards an articulatory goal, but how fast a certain articulator reaches its target depends on its spatial dis-

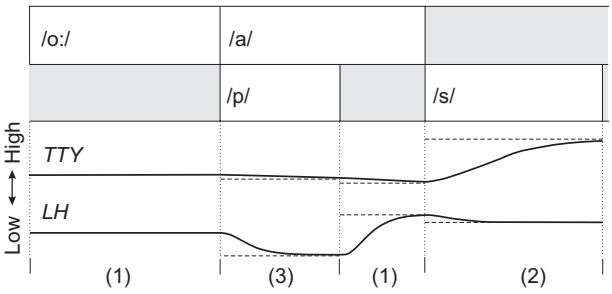

Figure 7: Time schedule for vowels and consonants for the utterance /opas/. The vertical dotted lines mark the boundaries of the composite motor commands. The two curves at the bottom show the resulting movement traces for the vocal tract parameters *TTY* (vertical tongue tip position) and *LH* (lip opening). The motor command targets for these parameters are depicted as dashed lines.

tance to the target and its relative speed. In the following, such a synchronized set of simple motor commands (one for each vocal tract parameter) defined over the same time interval will be referred to as *composite* motor command.

The high level input to the proposed control model is a time schedule for phonemes as illustrated in the upper two rows of Fig. 7. This schedule defines the "activation intervals" for the vowels (first row) and the consonants (second row) of an utterance. For the shown example, the phoneme intervals were adjusted according to the coordination principles suggested in [9]. This time schedule is transformed into a sequence of composite motor commands. Therefore, each phoneme boundary in either of the two rows is also a boundary in the generated sequence of composite motor commands. Each composite motor command represents either the target for a vowel (1), the target for a neutral consonant (2), or the target for a consonant overlapping a vowel (3). The cases are marked accordingly in Fig. 7. In the first two cases, the targets of the (simple) motor commands are given by the predetermined neutral target configurations for vowels and consonants. In the third case, the targets are determined by the dominance model described in Sec. 3.1. As additional input to the control model, the "articulatory effort" must be specified for each phoneme. This parameter is transformed into time constants for the simple motor commands taking into account the relative speed of the corresponding articulators. To illustrate the above concepts, the generated simple target sequences for the lip opening (*LH*) and the vertical tongue tip position (*TTY*) are shown at the bottom of Fig. 7 for the utterance /opas/. For this example, the time constant τ for all phonemes was set to 32 ms.

4. Discussion

We have introduced a new quantitative control model for an articulatory speech synthesizer based on a model for coarticulation and the concept of target approximation (TA). Originally, the TA concept was devised to explain the mechanisms underlying F_0 production [8], but in recent experiments by Xu and Liu [9], it became apparent that also the movement of the other speech articulators can be explained as a process of sequential target position approximation. The presented model is the first quantitative implementation of this idea in terms of motor commands for supraglottal articulators.

A comprehensive evaluation of the model remains to be done in a next step. One conceivable approach for an evaluation would be to test how well the control model can replicate articulatory movement traces of a natural speaker. Furthermore, the quality of synthetic speech generated using this control model should be assessed. For the latter case, informal synthesis and listening tests have already been performed by means of few example words. For that purpose, we additionally implemented the control of laryngeal parameters (i.e., glottal abduction, F_0, subglottal pressure) of a model of the vocal folds, also based on target approximation. The results were very promising and indicate that this control model can lead to a major improvement in the quality of articulatory speech synthesis.

5. Acknowledgments

This research was funded by grant no. JA 1476/1-1 from the German Research Foundation. The author would like to thank Yi Xu and Bernd J. Kröger for their comments and contributions to this manuscript and Sascha Fagel for making his EMA data available to us.

6. References

[1] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 2006, pp. 873–876.

[2] P. Birkholz and B. J. Kröger, "Vocal tract model adaptation using magnetic resonance imaging," in *7th International Seminar on Speech Production (ISSP'06)*, Ubatuba, Brazil, 2006, pp. 493–500.

[3] P. Birkholz, D. Jackèl, and B. J. Kröger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1218–1226, 2007.

[4] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[5] B. J. Kröger, *Ein phonetisches Modell der Sprachproduktion*. Niemeyer, Tübingen, 1998.

[6] E. L. Saltzman and K. G. Munhall, "A dynamic approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, pp. 333–382, 1989.

[7] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[8] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, pp. 319–337, 2001.

[9] Y. Xu and F. Liu, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics (in press)*, 2007.

[10] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1878–1889, 1995.

[11] K. Ogata and Y. Sonoda, "Evaluation of articulatory dynamics and timing based on cascaded first-order systems," in *Proceedings of the 5th Seminar on Speech Production*, Kloster Seeon, Germany, 2000, pp. 321–324.