Heikki Rasilo

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
The Artificial Intelligence Laboratory, Vrije Universiteit Brussel

heikki.rasilo@aalto.fi

Articulatory model for synthesizing sequences of arbitrary speech sounds or pre-programmed Finnish phonemes

Abstract: This document reports the characteristics and functioning of an articulatory model created in order to produce speech sounds based on articulatory gestures that capture the most important voicing, timing and articulatory configuration characteristics but can be arbitrary in nature. The model can be used to produce child-like babble by randomizing model parameters, when the language-dependent phonemes are not known, or to produce language-dependent phonemes from pre-defined phoneme characteristics. Finnish phonemes have been defined into the system, and Finnish speech can be synthesized from written input strings. Synthesis is based on articulatory target points, ATP's, in 9-dimensional articulatory domain, with corresponding timing and voicing characteristics. The point-to-point movement dynamics are assumed to be dependent on the physiology of the system and thus innate. Coarticulation is taken into account by using a look-ahead property when approaching following phonemes. Voiced and unvoiced consonants, fricatives, liquids, nasals and vowels can be synthesized with the model and animations of the mid-sagittal vocal tract movement can be created. The model has been used in simulations of infants' speech acquisition, where the caregiver uses pre-defined Finnish phonemes and the infant tries to learn the same phoneme system with different vocal-tract length and fundamental frequency, and without knowledge of the phoneme characteristics.

Table of Contents

1 Introduction

Vocal tract models have been used for decades when modeling several kinds of phenomena related to human speech. Several models for human articulation have been created for speech synthesis and analysis purposes since an efficient method for compressing speech became crucial for telecommunication purposes (e.g. [1], [2], [3], [4] and [5]). Acquisition of articulatory motor skills have been modeled using articulatory models ([6], [7]), researchers of evolution of speech have attempted to reconstruct vocal tracts of different species to investigate their sound repertoires ([8], [9]) and evolution of vowel systems in populations have been simulated using articulatory models [10]. Speech inversion aims to map acoustic speech into parameters of articulatory models ([11], [12], [13], [14], [15], [16]), and finally infants' speech acquisition has been investigated using articulatory models ([17], [18], [19]).

The development of our vocal tract model started while experimenting with techniques for speech inversion ([16], [20]). Details were added to the model gradually when increased accuracy in modeling certain properties of human vocal tracts became crucial for simulation purposes. When the focus shifted from language independent and mathematical speech inversion mappings towards a more plausible way to approach the problem based on natural phenomena present in infants' language acquisition situation, articulatory movement dynamics were added and the model was made able to babble arbitrary speech sounds like an infant or to produce language-dependent, learned, speech phonemes like a caregiver.

This document describes the vocal tract model used in our experiments of infants' speech acquisition, and aims to do it in enough detail so that the model can be accurately reimplemented. The general outline and principles are adapted from Mermelstein's articulatory model [3] but since the sizes of the vocal tracts and the articulators vary depending on the individual, and even same phonemes may be pronounced differently by different speakers (e.g. [21]), I have not intended to follow the description of the Mermelstein's model meticulously. The aim has rather been to create a simple model able to produce all Finnish phonemes with pleasing quality while using a minimal amount of adjustable parameters. The most important morphological structures of the vocal tract are modeled, and model parameters have been adjusted during the developmental process to obtain good synthesis quality and reasonable vocal tract area functions as well as formant frequencies. MRI data considering different phonemes by Story et al. [22] was used as a guide in order to obtain realistic mid-sagittal images and area functions of the vocal tract. The ability to produce Finnish vowel sounds was confirmed by continuous listening of the resulting synthesis by the author, as well as comparing the formant data into known formant frequencies of Finnish vowel sounds as reported by Wiik [23].

2 Obtaining static vocal tract area functions from a set of positions of fundamental articulators

This chapter describes how static vocal tract area functions are obtained from nine parameters related to the positions of fundamental articulators such as tongue base, lips and velum. The algorithm takes the nine parameter values as inputs and returns a vocal tract area function related to the parameter values, calculated according to the geometry the vocal tract model.

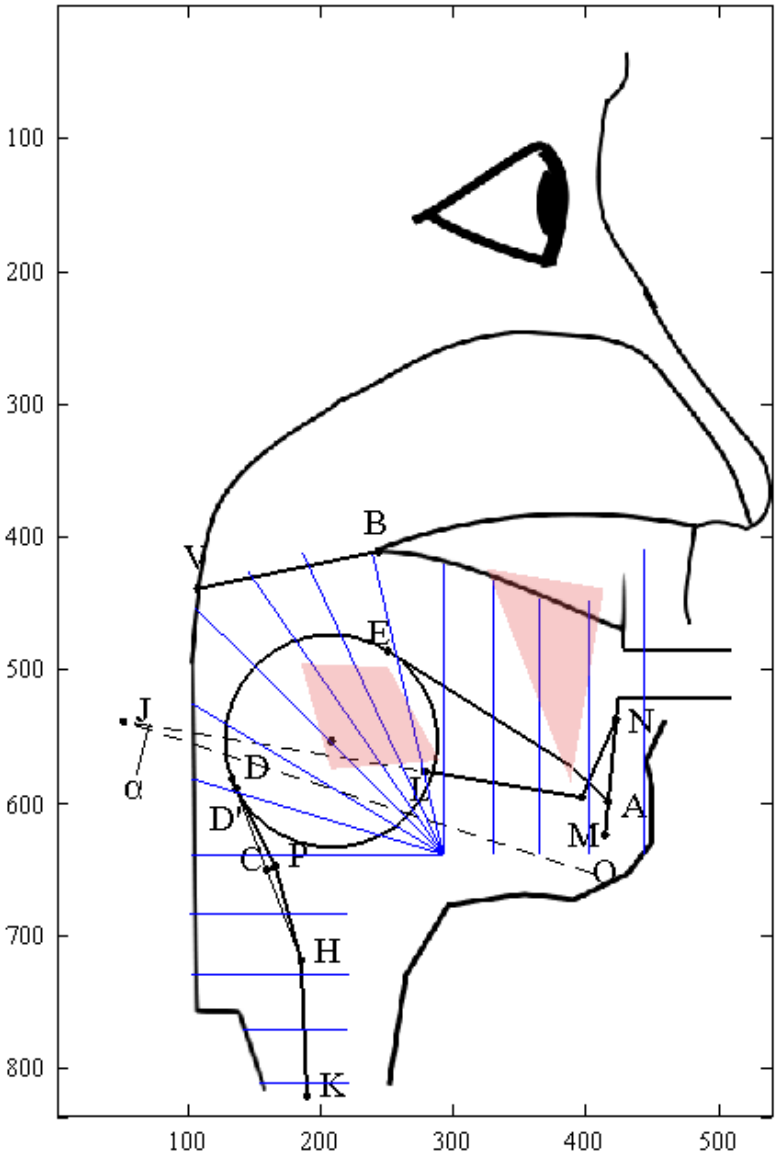

Figure 1. Image of the vocal tract model producing a Finnish vowel sound /o/. The blue lines indicate the positions of the 16 grid lines, at which the diameters of the tube segments are calculated. The nasal tract is modeled with constant tube areas, except for the first nasal opening tube (see the text). The light red areas show the ranges where tongue tip and tongue body parameters can get their values. The coordinate values correspond to pixels so that 1 cm equals about 37 pixels.

The outline of the vocal tract was originally sketched with image processing software, and for that reason I will list all the needed coordinate values in integers that represent the pixel positions in the image. The scale of the image is such that about 37 pixel units correspond to one centimeter in distance. The y-coordinate runs from the top of the image down, and the x-coordinate from left to right.

The cross-sectional area of the vocal tract is estimated at 16 locations indicated by the straight grid lines in Figure 1, and the obtained areas are used as the cross-sectional areas of 16 uniform co-axial tube segments in a Kelly-Lochbaum [1] tube model for the vocal tract. The cross-sectional surfaces at the first 15 positions are approximated as circles whose diameters $d_1 \dots d_{15}$ are defined by the lengths of the 15 lines limited between their intersections with the posterior-superior and anterior-inferior vocal tract walls. The exact ending coordinates of the lines are listed in Table 1. The 16th diameter d_{16} is estimated directly from the defined lip opening parameter as discussed in section 2.1.7. The resulting 16 areas are scaled with heuristically found scaling factors (Table 2) to compensate for the missing information about the third dimension of the vocal tract.

The posterior-superior wall is traced from the original vocal tract outline in [3], except for the velum (or soft palate) which is estimated as a straight line starting approximately at the highest point of the oral tract and connecting to the posterior vocal tract wall at point V in non-nasalized speech sounds. Except for the velum affecting sections 8, 9 and 10 in the tube model, the posterior-superior vocal-tract wall is considered stationary.

Table 1. Coordinates of the end points of the 16 grid line segments dividing the vocal tract into 16 cross-sections.

Section number	**Coordinates at the posterior-superior side (x,y)**	**Coordinates at the anterior-inferior side (x,y)**
1	(153, 812)	(220, 812)
2	(140, 772)	(220, 772)
3	(101, 730)	(220, 730)
4	(101, 685)	(220, 685)
5	(101, 640)	(291, 640)
6	(101, 583)	(291, 640)
7	(101, 526)	(291, 640)
8	(104, 454)	(291, 640)
9	(144, 427)	(291, 640)
10	(185, 413)	(291, 640)
11	(238, 415)	(291, 640)
12	(291, 422)	(291, 640)
13	(329, 434)	(329, 640)
14	(364, 448)	(364, 640)
15	(401, 450)	(401, 640)
16	(442, 411)	(442, 640)

Table 2. Scaling factors of the 16 cross sectional areas to compensate the variation in the third dimension.

Section number	**Scaling factor**
1	0.98
2	1.06
3	0.60
4	0.60
5	0.65
6	1.04
7	1.43
8	0.90
9	1.60
10	2.40
11	2.42
12	2.67
13	2.31
14	2.30
15	1.05
16	2.92

2.1 Parameters defining the vocal tract air space

This section describes the parameters and calculations needed to obtain an area function from the given parameter values. Nine parameters are given as an input to the vocal tract model, and their parameter values can lie inside pre-defined ranges. The nine parameters, their symbols in the following description and ranges are given in Table 3.

It has to be emphasized that the calculations based on the nine parameter values lead into a *static* shape of the vocal tract. In our simulations these parameters alone are not enough to define a *phoneme* or a *target articulatory gesture* of a given language. The phoneme is defined with additional parameters that have to do with the dynamics of the articulators as well as timing and voicing parameters, explained in more detail in Chapter 3. The calculations based on the nine vocal tract parameters providing the static articulatory shape can be considered as the *transformation of the positions of the articulators into a vocal tract area function of one individual speaker*. This transformation is language-independent, whereas the extra parameters defining a phoneme are considered language-dependent.

Table 3. Nine parameters related to the fundamental articulators, whose values define the vocal tract area function after given as input to the geometrical vocal tract model

N	Description	Symbol	Range
1	Tongue body x-coordinate	b_x	[175,290]
2	Tongue body y-coordinate	b_y	[467,550]
3	Tongue tip x-coordinate	t_x	[324,412]
4	Tongue tip y-coordinate	t_y	[425,530]
5	Jaw angle	α	[0,15]
6	Hyoid x-coordinate	H_x	[145,206]
7	Lip protrusion	l_p	[0,2]
8	Lip opening	l_o	[0,40]
9	Velum opening	V	[0,1]

2.1.1 Larynx

The horizontal position of the larynx *K*, moves in relation to the position of the hyoid *H*. According to Mermelstein [3], "*anterior-posterior movement of K is found to be one-half that of H*". Adapting similar policy the larynx *K* moves according to the formula 2 in Table 4. In our model the movement is one third of that of the hyoid, resulting in a final range of areas between 0.14 and 1.09 cm^2 for the first section. The neutral positions for the two variables were found empirically.

It was discovered that with wider larynx values, the highest frequency values of about 2500 Hz for the second formant could not be reached. According to Wiik [23] a long /i/ vowel in Finnish had a F2 value of 2495 Hz based on the mean of five male speakers. Ericsdotter [24] reports cross sectional areas between 0.41 cm^2 (in /i1/ and /y1/) and 1.39 cm^2 /ae1/ for a male subject's pharynx, but the acoustical measurements indicate an F2 value of less than 2250 Hz for /i/. Similarly Story [22] reports an area of 0.33 cm^2 for /i/ at the glottal end and the corresponding F2 value of 2332 Hz. The lower limit for larynx position was chosen lower than the reported ones in order to obtain F2 values of about 2500 Hz.

Different formulations and values for the behavior of *H* and *K* were experimented with, and the one with the most pleasing sound synthesis quality was chosen. For example, using constant larynx area (*K*=172*)* resulted also in good synthesis quality with recognizable phonemes, but the slightly varying position of *K* was found more comfortable. Variation in these particular parameters varies strongly the voice color of the synthesis; wider larynx results into more resonant voice, whereas tighter larynx gives a tense impression of the voice. Intelligibility of synthesized speech is not noticeably affected.

In Mermelstein's model *K* is located 2.7 cm below *H*, in our model the distance between the two is 102 units. The vocal tract wall between *K* and *H* is estimated in our model with a straight line segment.

2.1.2 Pharynx

Following Mermelstein's model the anterior outline of the pharynx is shaped by the point *P*, whose position is dependent on the distance *d* between the hyoid and tongue base. The distance *d* is calculated as the length of *HD*, *HD* being a tangent for the tongue body circle at the posterior side. The point P is found on the normal bisector of *HD*, at distance defined by formula $p = 0.57 \times \left(\frac{d}{37} - 3.48\right) \times 37$, where the extra division and multiplication by 37 (when compared to Mermelstein's formula in Figure 2 [3]) is needed to scale the used length unit into centimeters and back. The corrected outline is thus formed by lines *HP* and *PD*', where *PD*' is the tangent of the tongue body in the posterior side connected to point *P*. If at some occasion the point *P* is detected inside the tongue body circle, the line *PD*' is ineffective since the outline in this case is limited by *HP* and the tongue body.

2.1.3 Jaw

The jaw is estimated to rotate clockwise according to the point *J*. The rotation angle α is a controllable parameter. The points *L*, *M*, *N*, *O* and *A* representing the lower jaw, lower incisors and tongue ending point are in their neutral position when the jaw angle α is set to zero degrees, but are rotated about the point *J* according to α. The exact coordinates of the points in the neutral position are given in Table 4.

2.1.4 Tongue body

The tongue body is modeled as a circle with a radius of $r = 80$ units. Its center point's x and y coordinates b_x and b_y are controllable parameters whose final values depend on the jaw angle α. The final parameter values' dependence on the jaw angle enables more natural articulator movements when moving only the jaw moves also the articulators attached to it. The parameter values are given in the original coordinate system as if the jaw angle α was zero degrees and the final effective coordinate values are calculated automatically by rotating the given values clockwise about the point *J* with an angle α.

Considering the tongue body parameters, Figure 1 shows a polygon representing an experimental range of values that was found to be sufficient to produce all Finnish phoneme sounds. The range of values mentioned in Table 3 is a rectangle calculated from the extreme values of this polygon, and for computational simplicity the larger rectangular area is used in the experiments of Rasilo [19] to

randomize the infant's principal target coordinates for the tongue body. The coordinate values of the corners of the polygon are listed in Table 4. The polygon may be more convenient to use in babbling experiments, since it lacks the extreme and often physiologically implausible positions of the tongue body. In the experiments of [19], the virtual caregiver's tongue body target coordinates lay inside this polygon.

2.1.5 Tongue tip

The x and y coordinates of the tongue tip, t_x and t_y, are controllable parameters and like the tongue body, their final values also depend on the jaw angle. A few details of the behavior of the tongue tip have to be taken into account. First, the tongue tip has to be able to be in contact with the hard palate when the jaw is in down position as well as in top position (e.g. in utterances /ata/ and /iti/). Second, while the tip of the tongue is in contact with the hard palate, the jaw has to be able to lower without the position of the tongue tip changing (e.g. in utterance /ito/). Third, the tongue has to be able to rest at the down position and move in synchrony with the jaw movements (e.g. in /aka/). These effects are modeled by defining an area inside which the tongue tip can take coordinate values when the jaw angle is zero. In our model the area is triangular, with one corner at the resting position behind the lower incisors and two corners at the hard palate allowing for the constrictions in /t/ and English /r/. When the jaw angle changes, the down most corner of the triangle rotates about the point *J* changing the shape of the triangle. The tongue tip coordinate values, defined when the jaw angle is zero, are shifted to their final positions inside the triangle by an affine transformation **A**. The transformation maintains all target points inside the triangle and the uppermost corners of the triangle always at constant position. The tongue tip resting behind the lower incisors is rotated with the jaw and the effect of the rotation gradually degrades when approaching the upper locations in the triangular area.

The affine transformation **A** transforms the original triangle T_0 (when $\alpha = 0$) into another triangle T_α ($\alpha \neq 0$). The transformation is calculated using the corner points of the two triangles (see Table 4 for $\alpha = 0$). The down most point of T_α is calculated by the rotation of the down most point of T_0 about the point *J*. The solution for the affine transformation matrix can be calculated as follows:

$$
\begin{aligned}
&\mathbf{A}\begin{bmatrix} T_0^{x1} & T_0^{x2} & T_o^{x3} \\ T_0^{y1} & T_0^{y2} & T_o^{y3} \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} T_\alpha^{x1} & T_\alpha^{x2} & T_\alpha^{x3} \\ T_\alpha^{y1} & T_\alpha^{y2} & T_\alpha^{y3} \\ 1 & 1 & 1 \end{bmatrix} \\
&\Rightarrow \begin{bmatrix} T_0^{x1} & T_0^{x2} & T_0^{x3} \\ T_0^{y1} & T_0^{y2} & T_0^{y3} \\ 1 & 1 & 1 \end{bmatrix}^T \mathbf{A}^T = \begin{bmatrix} T_\alpha^{x1} & T_\alpha^{x2} & T_\alpha^{x3} \\ T_\alpha^{y1} & T_\alpha^{y2} & T_\alpha^{y3} \\ 1 & 1 & 1 \end{bmatrix}^T \\
&\Rightarrow \begin{bmatrix} T_0^{x1} & T_0^{y1} & 1 \\ T_0^{x2} & T_0^{y2} & 1 \\ T_0^{x3} & T_0^{y3} & 1 \end{bmatrix} \mathbf{A}^T = \begin{bmatrix} T_\alpha^{x1} & T_\alpha^{y1} & 1 \\ T_\alpha^{x2} & T_\alpha^{y2} & 1 \\ T_\alpha^{x3} & T_\alpha^{y3} & 1 \end{bmatrix} \qquad (1) \\
&\Rightarrow \mathbf{A}^T = \begin{bmatrix} T_0^{x1} & T_0^{y1} & 1 \\ T_0^{x2} & T_0^{y2} & 1 \\ T_0^{x3} & T_0^{y3} & 1 \end{bmatrix}^{-1} \begin{bmatrix} T_\alpha^{x1} & T_\alpha^{y1} & 1 \\ T_\alpha^{x2} & T_\alpha^{y2} & 1 \\ T_\alpha^{x3} & T_\alpha^{y3} & 1 \end{bmatrix}
\end{aligned}
$$

$$\Rightarrow \mathbf{A} = \left[\begin{bmatrix} \boldsymbol{T_0^{x1}} & \boldsymbol{T_0^{y1}} & \mathbf{1} \\ \boldsymbol{T_0^{x2}} & \boldsymbol{T_0^{y2}} & \mathbf{1} \\ \boldsymbol{T_0^{x3}} & \boldsymbol{T_0^{y3}} & \mathbf{1} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{T_\alpha^{x1}} & \boldsymbol{T_\alpha^{y1}} & \mathbf{1} \\ \boldsymbol{T_\alpha^{x2}} & \boldsymbol{T_\alpha^{y2}} & \mathbf{1} \\ \boldsymbol{T_\alpha^{x3}} & \boldsymbol{T_\alpha^{y3}} & \mathbf{1} \end{bmatrix} \right]^T$$

After obtaining **A**, the tongue tip coordinates are transformed into their final positions using:

$$\begin{bmatrix} \boldsymbol{t_x} \\ \boldsymbol{t_y} \\ \mathbf{1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} \boldsymbol{t_{x,\alpha=0}} \\ \boldsymbol{t_{y,\alpha=0}} \\ \mathbf{1} \end{bmatrix} \quad (2)$$

The tongue tip is connected with the point *E* so that the line *tE* forms a tangent for the tongue body circle. The tongue tip is also connected with point *A* to close the inferior outline for the 15th section line, and to avoid abrupt changes in the area of the 15th tube section when *t* crosses the 15th grid line.

The triangular range for the tongue tip is extended above the hard palate so that complete closure of the vocal tract can be modeled. If no intersection of the line segment *tE* and grid lines 13 or 14, or line segment *tA* and grid lines 13 or 14 is found, the tongue tip has reached complete closure with the hard palate and the areas of the corresponding tube segments are set to zero. Similarly, if $t_x \geq 401$ and $t_y \leq 450$, the numerical values referring to the x and y coordinates of the superior point of the 15th grid line, the area of the 15th section is set to zero.

2.1.6 Velum

The velum is connected to a stationary point *B* approximately at the highest point of the oral tract. When the velum is closed, the coordinates of V_0^x and V_0^y are listed in Table 4. The line segment *BV* is rotated about the point *P* counterclockwise in the angle $\beta = 20° \times v$ from its closed position, where *v* is the velum opening parameter in range [0, 1]. The maximum opening of the velum is thus 20 degrees. The diameters of the tube sections 9-10 are always affected by the intersections of the velum and the corresponding grid lines. The diameter of the 8th tube section is constrained either by velum or the original posterior coordinates of the 8th gridline, depending on the degree of velum opening. If the tongue body circle is detected to cross the line *VB* on the grid lines 9, 10 or 11, the corresponding section areas are set to zero to indicate complete closure of the oral tract. The area of the first section of the nasal tract is not estimated geometrically from the model but is defined directly from the parameter value of *v*.

2.1.7 Lips

The lip opening and protrusion are currently not estimated from the geometrical model, but are calculated directly from the input parameters. The diameter of the 16th section is twice the given lip opening parameter ($d_{16} = 2l_o$). The lip protrusion parameter l_p gets values in the range [0, 2] correspondingly scaling the length of the lip section linearly from one to two times its original length. The value of l_p corresponds directly to the coefficient *D* used in the Lagrange interpolator to calculate the amplitude of the back propagating wave at the junction of the 15th section and the lengthened lip section. More details on the calculation can be found in [20].

2.2 Nasal tract

The nasal tract is coupled between the 7th and 8th tube section of the oral tract using a junction of three acoustic tubes (see e.g. [25] for equations). The first section of the nasal tract has the area of the ninth parameter *v* in cm^2, corresponding to the velum opening. The nasal tract has 11 sections (a length of about 12 cm with the sampling frequency of 16000 Hz), whose other values than the first one stay constant: $A_{nasal} = \{v, 2, 4, 4, 5, 6, 6, 6, 4, 2, 1\}\ cm^2$. The shape and the length of the nasal tract are estimated roughly using [26] and [27] and adjusted when listening to synthesized nasal speech sounds.

2.3 Final area function and parameter values

The final areas of the 16 tube sections are calculated as

$$a_i = \pi \cdot (d_i/2)^2 \cdot s_i,\ \ i = 1, 2, \dots, 16 \qquad (1)$$

Table 4. Variables needed in the calculation of area function in our geometrical vocal tract model

Neutral horizontal position of the interior wall of the larynx	$K_{neut} = 186$
Neutral horizontal position of hyoid	$H_{neut} = 177$
Vertical position of the hyoid *H*	$H_y = 720$
Vertical position of *K*	$K_y = 822$
Horizontal position of *K*	$K_x = K_{neut} + (H_x - H_{neut})/3$ (2)
Jaw rotation point *J*	$\begin{cases} J_x = 50 \\ J_y = 540 \end{cases}$
Lower jaw and incisors	$\begin{cases} L_x = 280 \\ L_y = 540 \end{cases}\ \begin{cases} M_x = 400 \\ M_y = 540 \end{cases}\ \begin{cases} N_x = 416 \\ N_y = 478 \end{cases}\ \begin{cases} O_x = 422 \\ O_y = 565 \end{cases}$
Coordinates of the corner points of the tongue tip range when $\alpha = 0°$	$\boldsymbol{T_o^{x1}} = 390\quad \boldsymbol{T_o^{y1}} = 530$ $\boldsymbol{T_o^{x2}} = 412\quad \boldsymbol{T_o^{y3}} = 472$ $\boldsymbol{T_o^{x3}} = 324\quad \boldsymbol{T_o^{y3}} = 425$
Coordinates of the corner points of the tongue body range polygon (x,y)	(240,467), (175,475), (210,550), (290,530)
Tongue ending point	$\begin{cases} A_x = 420 \\ A_y = 540 \end{cases}$
Anterior velum attachment point *B*	$\begin{cases} B_x = 242 \\ B_y = 412 \end{cases}$
Posterior velum point *V* when velum opening is 0	$\begin{cases} V_0^x = 106 \\ V_0^y = 440 \end{cases}$
Diameter of the lip section	$d_{16} = 2l_o$

When parameters corresponding to eight fundamental articulators (velum is kept closed) are uniformly varied in their defined regions (tongue-tip and tongue-body values from the polygons in Table 4), area functions calculated and two first formant frequencies estimated from the impulse responses, the well-known vowel triangle can be plotted. Figure 2 shows the vowel triangle formed by our model with a vocal tract length of 17.5 cm. The red circles are reference Finnish vowel values from [23] and the green diamonds mark the vowels obtained by our model with the

parameter values listed in Table 6. The match between the formant values is reasonably good – in F1-F2 sense, exact matches for the reference values could be created, but they would not sound like pure prototypical vowel sounds. This can be explained by the fact that the perception of single vowel sounds is bound to the auditory context where it is presented [see e.g. 28]. The morphology of our model differs from the average vocal tract that pronounced the reference values in [23], and the modeled vowel prototypes sound the most natural for this particular speaker (i.e. the vocal tract model).

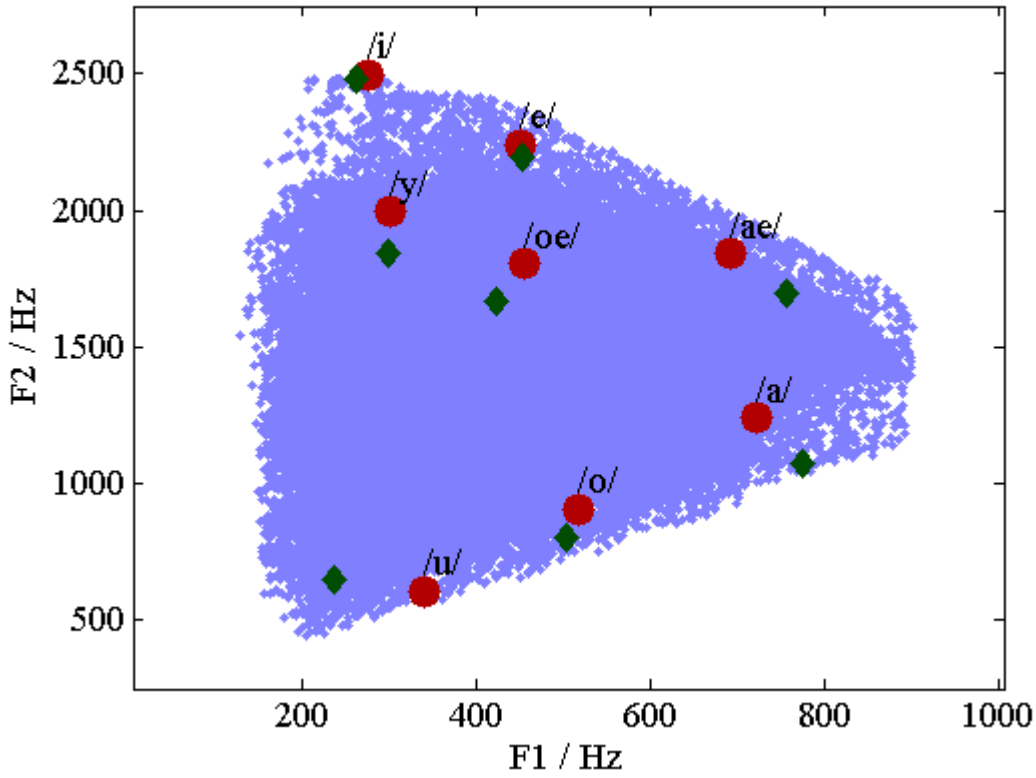

Figure 2. Scatterplot of the two first formant frequencies of about 200,000 unnasalized vowel sounds produced by the articulatory model. Sampling is performed uniformly through possible articulatory configurations. The well-known *vowel-triangle* is formed. The red circles mark the F1-F2 values of the reference Finnish vowels from the data by Wiik [23] and the green diamonds mark the values for the vowels defined in out model.

3 Dynamics of the vocal tract

The previous section described how a given set of parameter values corresponding to the positions of the fundamental articulators is transformed into vocal tract area function using our geometrical vocal tract model. This section describes how the dynamics of the vocal tract is designed, i.e. how the nine vocal tract parameters are smoothly varied during speech synthesis resulting into realistic synthesis quality. Smooth trajectories for the nine parameter values are calculated, and these parameter trajectories are later input in the geometrical vocal tract model to obtain trajectories for

the area function. These area function trajectories are in the last phase synthesized into sound. The synthesis phase is described in chapter 4.

Speech is often described as a generatively created sequence of a limited set of articulatory events, allowing the production of a very large (or infinite) number of different utterances. (e.g. [29]). Throughout this report, I equate the term *phoneme* with a *target articulatory gesture* (or event). The phoneme is thus a gesture leading into reaching and releasing its *articulatory target position* (from this on, ATP) with proper timing and voicing parameters. Importantly, the nine articulatory parameters explained in chapter 2 define an ATP, but are not enough to define the complete characteristics of a phoneme. The phoneme is a language-dependent unit that a speaker of a language has to learn.

It was noticed after the implementation of our vocal tract model that the rules controlling the dynamics of the vocal tract are to some extent related to the look-ahead-model implemented by Henke [2]. The actual differences of the two models are outside the scope of this documentation, and only the functioning of our model is discussed in detail. The author is aware of the existence of vocal tract models that are able to explain many speech related effects including motor equivalence, coarticulation and speech rate effects (e.g. task-dynamic model of Saltzman and Munhall [6] and DIVA by Guenther [7] but since accurate modeling of these effects is not in the focus of our current research, they were not intended to be replicated in the current version. Including the more realistic control structures in our model will be considered in future simulations. Our model's articulatory control uses points in the articulatory domain as the ATPs (as opposed to e.g. convex regions by Guenther [7]) but achieves reasonable synthesis quality and coarticulation effects for the purpose of our study.

We make a distinction between the properties of consonant and vowel phonemes. Vowels are more vaguely defined, and their ATPs do not have to be exactly reached in order to produce intelligible speech. Coarticulation is known to change the properties of phonemes based on the surrounding phonemes. Especially, *vowel reduction* [30] or *target undershoot* refers to the shift of the vowel formant frequencies towards the adjacent consonantal sounds in CVC contexts.

As opposed to vowels, consonants' ATPs are defined as a subset of the nine parameters, and in our model, they have to be reached exactly as they are defined (except if another consonant affecting the same parameters follows, see the detailed description later), and they thus have a *priority* over vowel targets. Intuitively, if for example the ATP for closure /t/ is not completely reached, the vocal tract remains partly open producing /s/, or some vowel sound leading to severe problems in speech intelligibility. In reality, coarticulation affects also consonantal sounds. For example, Öhman [31] has shown that the exact location of the velar closure during /g/ in utterances /ygy/ and /ugu/ is somewhat different. This effect is captured in Saltzman's and Munhall's [6] model as well as in DIVA [7], but not in the current version of our vocal tract model.

The dynamics of the nine vocal tract parameters are maintained smooth using minimum-jerk trajectories. Flash & Hogan [32] have shown that human arm's point-to-point reaching movements are approximately straight with a bell-shaped tangential velocity curve. The best function to describe the movement is a minimum-jerk trajectory, meaning that the derivative of the acceleration

of the movement is minimal when integrated over the entire movement. We make an assumption that the movements of the articulatory parameters are planned in a similar fashion.

3.1 Parameters defining the phonemes

This section describes all the properties that define a *phoneme* in our model. In addition to the nine parameters defining the ATP, seven parameters are needed to define the timing, voicing and priority characteristics related to a phoneme. The seven parameters are chosen in order to capture the most important characteristics of phonemes. Most importantly, approach and release durations aim to capture the effects of the velocities of the articulators, voice onset time (V_{on}) is known to depend on the closure position and the language in question and closure duration is found to depend on the closure position (see e.g. [33] and references therewithin). All these properties are free parameters to be defined in our model in order to achieve intelligible language-dependent phoneme sounds. All the phoneme dependent parameters are described in Table 5.

Table 5. List of all properties needed to define a phoneme.

Property of the phoneme	Definition	Symbol
Articulatory target position (ATP)	The nine vocal tract parameters defining the target position of the phoneme in the articulatory domain.	See Table 3
Approach duration (or target lookahead-time)	Defines, how long before the target instant the approach towards it is started.	D_a
Hold duration (or closure duration for consonants)	Duration of holding the articulators in the target position when achieved.	D_h
Release duration	Duration of the target release towards the next target after the hold period.	D_r
Voice onset time, V_{on} (Only for consonants)	The duration between target release and voicing.	V_{on}
Voice offset time, V_{off} (Only for consonants)	Defines, how long before the target voicing ends.	V_{off}
Excitation	Excitation during the target, 0 = silence, 1 = voicing, 2 = hiss (used in /h/).	E
Priority (consonant / vowel)	Defines if the target overrules the current target (generally, consonants replace all targets, but vowels do not replace consonants).	P

The timing parameter values for Finnish phonemes have been finally chosen based on unreported phoneme quality comparison tests by the author when varying the parameter values. A few guidelines have been followed when choosing the values for Finnish phonemes:

- In the process of creating the articulatory synthesizer for the caregiver's phonemes, we noticed that in order to perceive correct phonemes in the case of consonants it is important that the approach and release durations are longer for velar consonants than alveolar or labial consonants. This has to do with the fact that the velocities of the articulators depend on their masses so that tongue base is the slowest to move, tongue tip the fastest and lips there in between ([34] cited in [33]).
- The voice onset time (VOT) is reported to grow the further back the place of articulation is located (see [33] with several citations). Suomi [35] has reported VOT times of 11, 16 and 25 ms for /p/, /t/ and /k/ correspondingly. We ended up in VOT times of 20, 20, and 30 ms correspondingly.

- Maddieson ([36], cited in [33]) reports that the behavior of the hold time (in case of consonants same as closure duration) is reverse to the VOT so that the further back the place of articulation, the shorter the closure duration. This policy has been replicated in our model with unvoiced stops, /p/, /t/ and /k/ having hold times of 60, 40 and 20 ms correspondingly.
- According to Ogden ([37], cited in [38]) closure duration for /d/ is much shorter than /t/. We ended up in a closure duration of 10 ms for /d/ and this resulted in notable increase in /t/-/d/ identification task.
- Brown & Koskinen [38] cite Lahti (1981, publication not mentioned) for reported average VOT time of 30.4 ms for /d/ and 12.4 ms for /t/. Our model uses 30 ms for /d/ but no advantage in identification was noticed when compared to 10 ms.

The timing parameters are illustrated in Figure 3, where the trajectory of the tongue body x-coordinate is drawn during an utterance /aka/.

The liquid /l/ is modeled by not letting the tongue tip reach the hard palate, simulating the effect of open cavities on the sides of the tongue in real human production. /s/ is modeled by defining the tongue tip coordinates so close to the hard palate that frication is created. This appears realistic, since also humans have to learn to move the tongue in such a position that the physical properties of the vocal tract create frication noise. /v/ and /f/ are created in a similar manner considering the lips. /v/ includes voiced excitation whereas /f/ does not. All the ATPs, timing and excitation properties for all Finnish phonemes used in our studies are listed in Table 6.

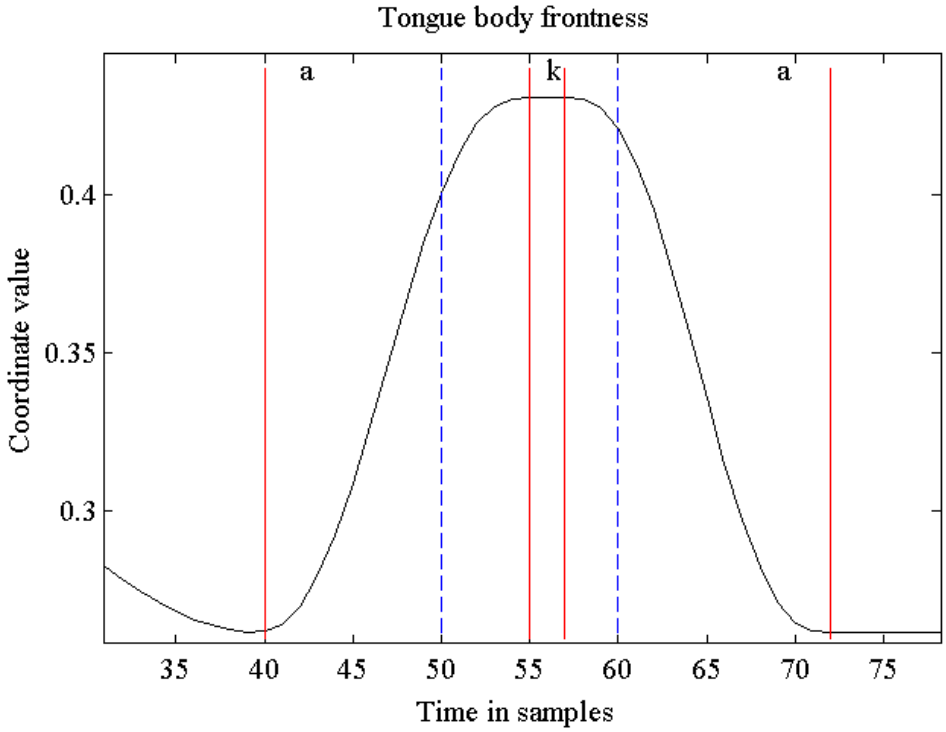

Figure 3. An example of a trajectory created for tongue body (x-coordinate) during the utterance /aka/. The ranges in between the red solid lines from left to right define the look-ahead time, hold-time and release time for the phoneme /k/. The blue dashed lines indicate the voice offset and onset instants correspondingly.

Table 6. All parameter values defining Finnish phonemes used in our study. Parameters marked with n are not affected by the corresponding phoneme.

	Articulatory target position (ATP)									Timing and excitation parameters						
Phon.	b_x	b_y	t_x	t_y	l_p	l_o	α	H_x	v	D_a	D_h	D_r	P	E	V_{on}	V_{off}
/a/	209	545	391	525	0.2	40	10	155	0	20	3	0	1	1	0	15
/b/	n	n	n	n	n	0	1	150	0	10	3	10	2	1	0	0
/c/	235	520	403	467	n	n	n	150	0	8	3	15	2	1	1	1
/d/	n	n	403	440	n	n	n	150	0	10	1	10	2	1	3	0
/e/	248	505	373	497	0	40	3.3	147	0	20	3	0	1	1	0	15
/f/	n	520	n	n	n	2	1	200	0	15	3	8	2	0	0	0
/g/	181	480	n	n	n	n	4	150	0	15	2	15	2	1	0	0
/h/	n	n	n	n	n	n	n	n	n	20	3	15	2	2	0	0
/i/	274	512	388	479	0	40	0.9	135	0	20	3	0	1	1	0	15
/j/	274	512	388	479	0	40	0.9	148	0	8	3	10	1	1	0	15
/k/	231	480	n	n	n	n	4	200	0	15	2	15	2	0	3	5
/l/	235	540	403	470	n	n	n	150	0	8	1	10	2	1	0	15
/m/	n	n	n	n	n	0	1	n	1	10	3	6	2	1	0	15
/n/	250	n	403	440	n	n	n	n	1	8	3	15	2	1	0	15
/o/	207	529	391	523	1.4	25	13	184	0	20	3	0	1	1	0	15
/p/	n	n	n	n	n	0	1	200	0	10	6	10	2	0	2	2
/r/	235	520	403	470	n	n	n	150	0	8	3	15	2	1	0	15
/s/	n	n	405	460	n	n	1	n	0	15	3	15	2	0	5	5
/t/	n	n	403	440	n	n	n	200	0	10	4	10	2	0	2	1
/u/	201	522	382	518	2	8	3	204	0	20	3	0	1	1	0	15
/v/	n	520	n	n	n	5	1	200	0	10	10	8	2	1	0	15
/y/	283	517	391	485	2	21	3	161	0	0	3	0	1	1	0	15
/z/	n	n	405	460	n	n	1	n	0	15	3	15	2	1	0	0
/ä/	241	522	387	504	0	40	8.6	150	0	20	3	0	1	1	0	15
/ö/	298	530	399	508	1.5	29	11	175	0	20	3	0	1	1	0	15
/ŋ/	221	470	n	n	n	n	4	145	1	10	3	6	2	1	0	15

3.2 Calculation of articulatory trajectories

The trajectory estimation algorithm calculates dynamic trajectories for the nine articulatory parameters defining the final shape of the vocal tract. The algorithm calculates trajectories from a string of phonemes input to the system in a form of a "timeline" or "phonetic score" – that is, the input consists of phonemes and their exact time instants defining when they are intended to be reached. The time instant given to a phoneme refers to the first time sample of the hold period (equals the first time instant of the closure period for consonants).

The use of a look-ahead parameter for vowels enables coarticulation. The model always tries to reach the *last* vowel in the defined *vowel-look-ahead* -region. This parameter is similar to the approach duration in the case of consonants. The look-ahead value for vowels is kept constant at 200 ms, i.e. the approach towards a vowel target is started 200 ms before the intended time instant for the target.

The trajectory estimation follows a few simple principles, which are assumed to be innate and originating from the physiology of the vocal apparatus. The minimum-jerk trajectory estimation takes as parameters the location, velocity and acceleration in the starting point and the hypothesized ending point of the articulatory parameter. Trajectory estimation begins using the neutral vocal tract configuration before the first target, and at every time step a minimum jerk trajectory is estimated to the selected next target point in the articulatory domain using current location, velocity and acceleration of the parameter. The velocity and the acceleration at the goal target are set to zero, i.e. the movement of the articulator is intended to stop at the target position. We use a time step Δt of 10 milliseconds, leading into an area function vector every 10 ms.

At each time instant t for each articulatory parameter p, the position, velocity and acceleration are calculated based solely on the position, velocity and acceleration of p at time instant t-1 and the goal phoneme's ATP value at the goal time instant t_g. Because of this, only rules for selecting the target phonemes and their time instants for each parameter are needed to guide the estimation. The timeline is gone through from left to right, and the phoneme target of a vocal tract parameter p (e.g. tongue-body), can change in the three cases of the following list. In the list, targets with priority are referred to as consonants, and targets with no priority as vowels:

1. A consonant appears so that its approach period starts at time t. As consonant overrules all targets, even if the parameter was performing another consonant gesture, the new one will replace the old target.
2. The current target is a consonant and it has been approached and held for the duration defined for the target - that is - release of the consonant has to be performed onto the next target. In this case the vowel-look-ahead region is gone through from the end towards the beginning (at time instants [*t, t + vowel-lookahead*]), and the last vowel target found from the region is chosen as the target where the consonant will be released. This vowel target will be assigned to all nine parameters, and the corresponding time instant t_g will be set in the end of the consonant's release period for all nine parameters. Thus the original intended time instant of the vowel will change according to the consonant that precedes it. This is important since the release time of the consonant seems to be an important acoustic cue in the consonant identification, and the release time is desired to stay constant. All parameters are affected in order to avoid "jitter" when some parameters would reach the target later than the parameters affected by the consonant. If no vowel target was found in the vowel-look-ahead region, the previous vowel target will be used.
3. The current target is a vowel and a new vowel appears on the distance defined by the vowel-look-ahead parameter.

These rules are used in order to enable the following properties

1. A new consonant always changes the target for the parameters it is defined to affect. For example in an utterance /asta/ the consonants /s/ and /t/ both affect the tongue tip, and in this case /s/ is not completely released to the following /a/, but approach towards /t/ is started instead.
2. Consonant is always released to some target during the consonant-dependent release period (except if a new consonant target appears which affects the same parameters, see point 1).
3. A new target may appear before the current vowel target is completely reached. This allows coarticulation. For example, in utterance /iai/ the target articulatory configuration of /a/

might not be completely reached (depending on the speech tempo) before starting the movement towards the final /i/.

The consonants are defined to affect only a limited set of the nine possible articulatory parameters. For example /t/ affects only the tongue tip's x and y -coordinates, hyoid and velum and /p/ affects lip aperture, jaw, hyoid and velum. Consonants are superposed on vowels only with respect to the parameters which are affected by the consonant. For example in an utterance /ati/ (see illustration in Figure 4), during /t/, there is a continuous trajectory from /a/ towards /i/ on the part of the 6 parameters that are not affected by /t/. This is considered as a characteristic of canonical babbling and thus innate.

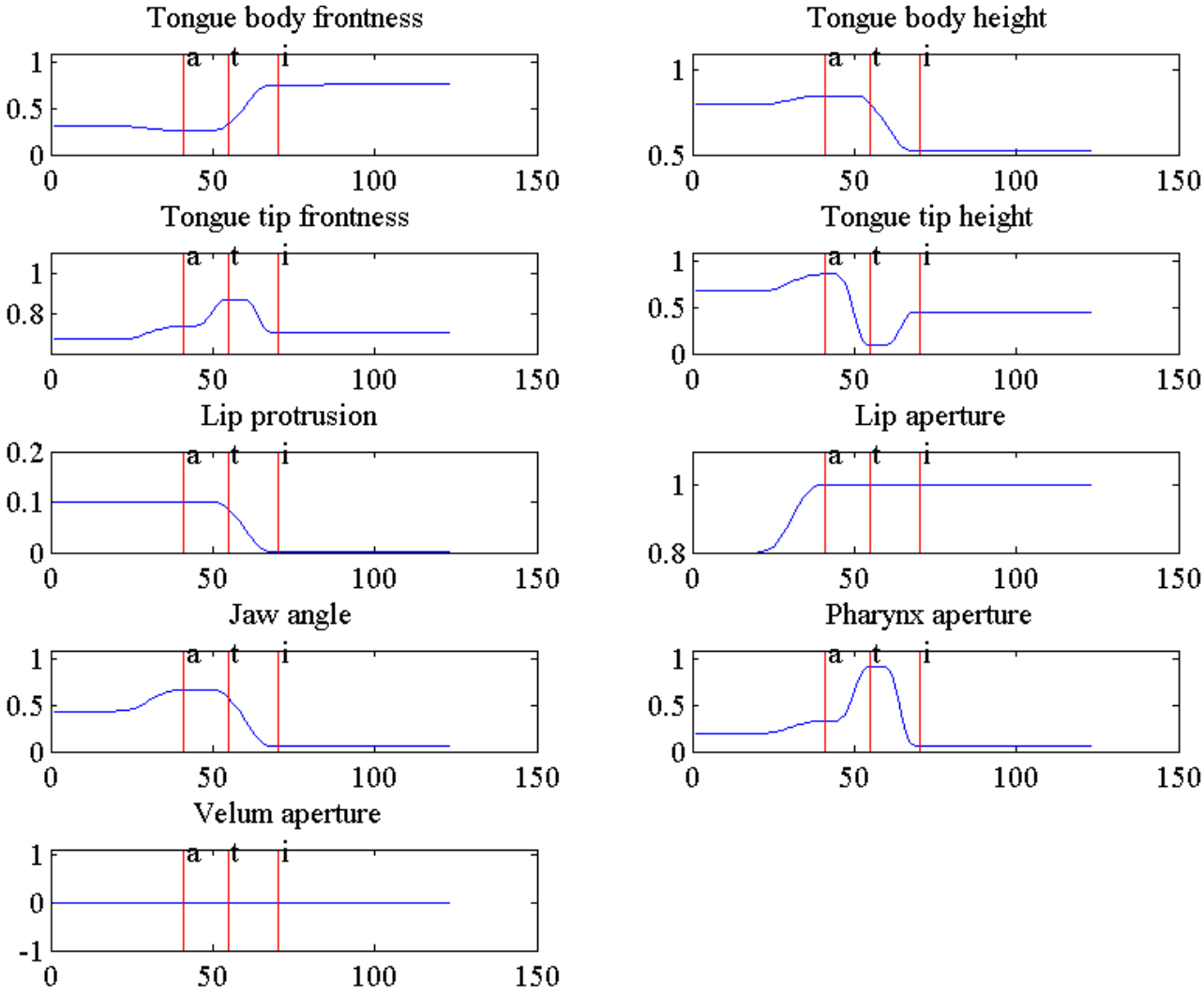

Figure 4. Articulatory synthesis of an utterance /ati/. It can be seen that only tongue tip and hyoid (and velum) parameters reach the target position for /t/ as is expected. Also it is seen that all the parameters are released to the target position of /i/ before the originally defined target time instant. This is because the release time defined for /t/ comes to an end before the target instant of /i/ and due to our assumption the consonant has to be released according to this release time. x-axis in the figures corresponds to time in samples and y-axis to coordinate value scaled into a range [0, 1].

It was also noticed that when a consonant is released, not only the parameters affecting the consonant but all parameters should arrive at the following vowel target at the same time instant in order to avoid "jitter" in speech synthesis, which is easiest noticed in utterances like /papipapipa/ (see also the second points in the previous lists). The /p/ has to be released in the following target with all parameters; otherwise the lips, jaw and velum would appear at the target in the end of the release time defined for /p/ and other parameters slightly later when the actual target position is reached. Currently we do not have physiological data verifying if such an effect happens in real speech but for our synthesis purposes it seems necessary, and in our simulations this property is considered innate. The mentioned effect can be only noticed in variegated babbling where the vowel sound varies during the babbled sequence. In our current simulations considering reduplicated babbling where the vowel sound does not vary during an utterance, the property is not crucial, but implemented for speech synthesis purposes.

The calculation of the trajectories should work with any chosen ATP-values and timing parameters, making it flexible to model other languages or for example babbling, where the language dependent constraints are not known.

3.3 Synthesizing speech from a string of phonemes

If speech is synthesized directly from a string of phonemes, our model puts the first target in a time instant defined by an offset which allows the articulation to start from the neutral position and reach the first target with its corresponding approach duration. Offset for our model is half a second. Phoneme targets are placed onto the timeline from the first to the last using constant spacing defined by parameter *target_spacing*, which tells the distance between the beginning of the release period of the current target and the beginning of the hold period of the following target.

In case of multiple repeated phonemes (e.g. a double consonant), a preliminary algorithm goes through the input string and lengthens the hold duration of the first token by h_{extra} for every repetition, which is adjusted to produce reasonably good sounding double consonants in Finnish. The values used in Finnish speech synthesis are shown in Table 7.

3.4 Effect of speech rate

Speech rate is refined by one parameter, *rate*, and affects parameters *target_spacing* and *h*extra. Vowel-lookahead is not affected, causing an effect where increased speech rate increases coarticulation. For a normal speech rate *rate* = 1 is used, for faster speech the value is decreased and for slower speech increased.

Table 7. Parameter values used in synthesizing Finnish speech from an input string of phonemes

Parameter	**Value**
target_spacing	$140\ \text{ms} \times rate$
h_{extra}	$250\ \text{ms} \times rate$
vowel-lookahead	200 ms
Δt	10 ms

3.5 Implementation of the dynamical vocal tract model

The details of the implementation of the vocal tract model may depend on the programmer, but a few details of my implementation are given here. I have used a timeline where at every intended target time instant (beginning of hold period), time instants of the beginning of the approach period, beginning of release period and end of the release period are also saved. This allows for effective searching of necessary time instants characteristic to a phoneme target. One variable keeps always track of the following consonant and the time instant of the beginning of its approach period for a parameter. This makes it easy to overrule other targets as soon as an approach period for a consonant is started. A pseudocode of the implementation is provided in Appendix A. The real implementation is done without the for-loop through the parameters p but all the variables are treated as vectors with a length of nine for faster calculation.

3.6 Conclusions from trajectory estimation

This section has described how a sequence of phonemes, intended to pronounce at certain time instants, is transformed into continuous trajectories of the nine parameters corresponding to our elemental articulators. Different velocities of individual articulators are modeled by using varying approach and release durations, coarticulation effect is taken into account in some extent in case of vowel sounds. The algorithm calculates trajectories with similar logics from any defined phonemes, and in principle it could be interpreted as an innate motor control system employed by all human speakers.

As usual, several simplifications have been made in the model construction, some control strategies characteristic to human speech have not been modeled (such as motor equivalence), and I cannot point out evidence that the rules used in the calculations are physically plausible, universal or language independent. Nevertheless, what is important is that the synthesis quality is comfortable enough to listen to, intelligible, and when listened, invokes a reaction of hearing a child speak. Context dependence of vowel and consonant synthesis also makes speech and phoneme recognition experiments more realistic and challenging when performed towards the synthesized speech. Since the calculation of the trajectories, be them realistic or not, is based on the characteristics of phonemes, speech acquisition mechanisms can still be investigated. For example, assuming a child that tries to produce native language–like phonemes, but does not master the correct characteristics used by the caregivers, produces less intelligible babble. When learning occurs, the phonemes should shift towards the targets known by the caregiver as in [19]. Different principles governing speech learning can be tested even if the underlying trajectories were not exactly the same as in real-life situations.

The trajectories of the nine parameters obtained for the phoneme string are input to the geometrical vocal tract introduced in Chapter 1. Smooth trajectories for the vocal tract area function are thus obtained, taking values every 10 milliseconds. The next chapter describes how these dynamical area functions are transformed into speech sound.

4 From area function trajectories into acoustic signal

This chapter explains, how an area function trajectory is synthesized into a speech signal. Modeling glottal excitation, vocal tract turbulences and wave propagation are dicussed. From the area functions, areas of less than 0.01 cm^2 are set to zero before this step.

4.1 Excitation type vector

First of all, a vector defining the type of glottal excitation, being as long as the final vector including the area functions to be synthesized, is created. Thus, every 10 milliseconds throughout the articulated utterance the glottis can take a new excitation value if needed. The *excitation type* vector can take three values: 0 for silence, 1 for voicing and 2 for hiss. In principle, voicing values should be defined by voice onset and voice offset parameters, but a few adjustments are needed.

1. In case of consecutive voiceless consonants, voicing in between is not allowed, even though it might be possible using only the voice onset and offset values. Thus segments of consecutive voiceless consonants are searched and excitation values between the voice offset time of the first consonant and the voice onset time of the last consonant are set to zero.
2. The same operation is performed for the excitation type of hiss (for example in a case a double /h/)
3. To model the rolling /r/, I have not modeled the rapid variations of the tongue tip yet, but the acoustic consequence is modeled by setting the excitation of every fourth value from the target time until the release time of /r/ to zero.
4. On time instants where vocal tract area function value is less than 0.01 cm^2 and the value of the velum parameter is less than 0.01 (otherwise, closed vocal tract may be nasalized), excitation is set to zero.

4.2 Fundamental frequency trajectory

The synthesis algorithm takes the desired fundamental frequency (F0) as an input (variable F_0), but certain variation to the F0 is applied with an *ad hoc* method to create more natural sounding utterances. A dynamic F0-trajectory, $f_{F0,}$ having an F0-value at every time instant of the vocal tract area function, is created. The starting F0 value, f_{start}, where the variation is started gets the value of F_0 plus a random integer drawn uniformly in the range of [1, 20]. From this value the fundamental frequency is accelerated with a random acceleration a_{rand} drawn uniformly in the range of [-120, 80] Hz/s^2 for a random time drawn uniformly in the range of [30, 120] ms. At every time step, the acceleration is resisted with a resistance acceleration $a_{resistance}$ towards f_{start} to avoid drifting too far from the starting fundamental frequency. The resistance is calculated using $a_{resistance} = 3 \cdot (F0(t-1) - f_{start})\ Hz/s^2$. This acceleration process is looped until values have been obtained for every time instant in the vocal tract area function vector. The result is a dynamically varying fundamental frequency curve that does not drift too far from the user-defined F0 for the synthesis. The variation in F0 not only creates a more natural sounding speech output, but also creates natural variation in the speech signals used in the language acquisition experiments so that the used speech signals are not exactly equal.

4.3 Glottal excitation signal

Synthesizing the area functions into a sound signal requires a *glottal excitation signal*, whose fundamental frequency should follow the trajectory created in the previous phase. The voiced segments of the excitation type vector are separated (selecting the segments that only consist of ones) and fed to an algorithm calculating the glottal excitation signal for these segments. Since the separation and the characteristics of the glottal pulses have to be related to the values present in the F0-trajectory vector, the corresponding measurements of different parts of the *glottal excitation cycle* are calculated according to the value of $F_0(t)$, where t corresponds to the time instant of the ending of the previous cycle, scaled to match the 10ms separation of samples in F_0. One *glottal excitation cycle* in our simulations is constructed from the following parts in exact order: opening of glottis, closing of glottis, glottal closure noise, glottal closure (silence), and glottal opening noise. The cycle is illustrated in Figure 5.

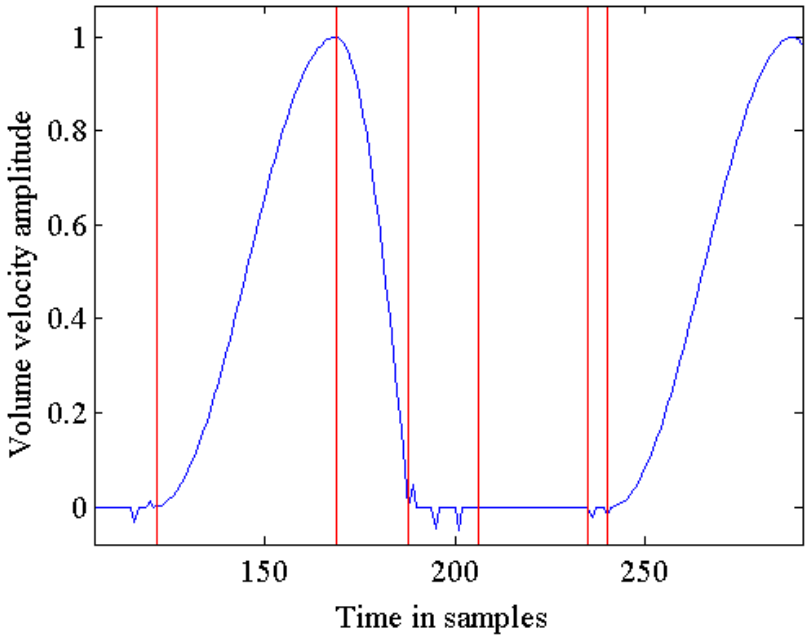

Figure 5. Glottal cycle. The vertical lines limit the phases of opening of glottis, closing of glottis, closure noise, glottal closure and opening noise (from left to right). After the last phase, new cycle is started with timing parameters calculated corresponding to the time instant of the end of the previous cycle in the fundamental frequency vector. The amplitudes of the noise segments are exaggerated in this image for clearness.

As the basis for the glottis pulse for volume velocity (or flow), pulse C in [39] is used. The values for T_P and T_N corresponding to the parts of the pulse with positive and negative slopes correspondingly are selected to be 40% and 16% of the cycle length. A few impulses of noise are added to the excitation signal before the glottis is opened and after the glottis is closed to slightly affect the color of the voice. The lengths of the noise vectors for opening and closing are $N_{open} = 15\%$ and $N_{close} = 4\%$ of the cycle length. In our current model, noise is added only after the glottal closure every six samples with a random amplitude uniformly drawn between [-0.005,

0.005] times the maximum amplitude of the glottal pulse. Finally, the glottal closure phase is modeled as a vector of zeros with a length of 25% of the cycle length

For every voiced segment, as determined from the excitation type vector, a whole number of glottal excitation pulses have to occur during the glottal excitation. In other words, if it is detected that voicing should be stopped in the middle of the opening-closing-period, the last pulse will not be generated.

The beginning of every voiced glottal excitation segment is set to rise smoothly to its maximum amplitude of one by using the raising half of a Hamming-window function of length of 20 milliseconds. Similarly, the endings of the segments are faded out by using the descending part of the same window function. The final pulse shape is lowpass filtered in order to cut out frequencies larger than half of the sampling frequency. Figure 6 shows a final filtered glottal excitation signal for a short vowel sound.

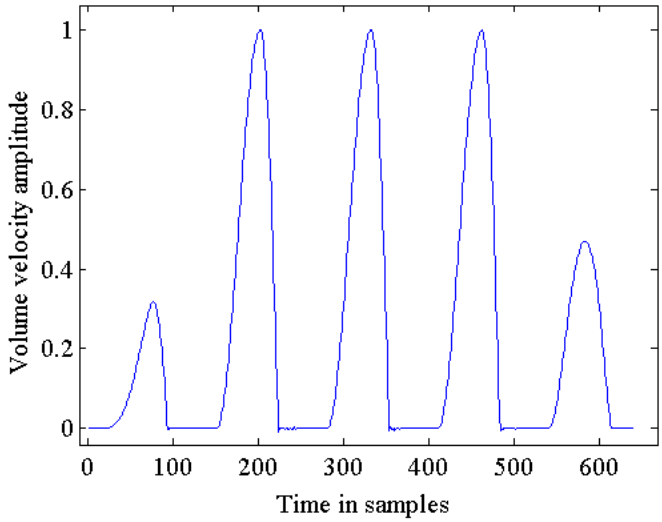

Figure 6. Final excitation signal for a short vowel sound.

The consonant /h/ is modeled by adding a hiss excitation to the glottis. Similarly to the voiced segments, segments that require hiss excitation are separated. These correspond to segments in the excitation type vector consisting of values two. The glottal excitation signal corresponding to these segments is set to uniform random noise drawn from range [-0.005, 0.005]. The beginnings of the noise segments are smoothened using Hamming window equally to the voiced segments.

4.4 Turbulence

Turbulences are known to occur in the vocal tract when a narrow vocal tract constriction forms a high-speed jet of air. Turbulent noise sources can be located at the constriction and the obstacles that the jet of air faces, for example the incisors [40]. In our model, fricative consonants are modeled with a random turbulent noise added downstream from the location of the constriction.

When any of the vocal tract tube segments has a small enough constriction for long enough time, turbulence noise is added to the corresponding location. The time constraint is used in order not to

add frication when short durations of narrow constrictions occur during the release of stop consonants. Since the area function changes every 10 milliseconds, noise would otherwise be added according to this duration, being too long for modeling the transients at consonant release. A transient and frication lasting for just a few milliseconds are known to occur at the consonant release [41].

Impulse-like turbulences with very short durations at the release of the consonant were also experimented with, but were not noticed to give improvement in consonant identification. A transient-like effect is created automatically due to the wave propagation calculations, when the glottal flow reaches the narrow opening in the vocal tract at the beginning of the consonant release, where the area of the constriction changes abruptly from zero into a small positive value.

Practically, frication noise in our model is added if the constriction area is between 0.01 cm^2 and 0.1 cm^2 for a duration of at least 40 ms. The maximum gain of the noise is inversely proportional to the area of the constriction so that between the two limits its gain g is linearly changed from 1 to 0.1. If the conditions are met, white uniform noise is added to every sample of the forward propagating volume velocity wave of the segment following the last segment that has the constriction, or to the last segment of the constriction if it is the last tube segment of the vocal tract. Amplitude of the noise is drawn from range [-0.03g, 0.03g].

4.5 Wave propagation in the vocal tract model

The sampling frequency used in our wave propagation calculations is 16,000 Hz, meaning that during every 10-millisecond static segment in the area function trajectory 160 wave propagation steps will be calculated. The forward and back propagating waves in the nasal and oral tract are attenuated by an area dependent loss factor at each step of the calculation at each tube segment. The loss factor α for section k is calculated using the same formula as in [42]:

$$\alpha_k = 1 - \frac{0.004\Delta x}{\sqrt{A_k}} \tag{3}$$

where A_k is the cross-sectional area of the tube segment, Δx is the length of the tube segment (1.1 cm in our case).

The 16th (i.e. lip-) section can be lengthened from one to two times its original length using a Lagrange interpolator to calculate a proper value for the back reflecting wave from the end of the 16th section from 5 saved consecutive forward propagating wave values at the lip section. Frequency dependent lip-radiation losses are taken into account with the method of Laine [43]. Equally, radiation losses are calculated for the ending of the nasal tract at the nostrils.

Wave propagation is calculated using Kelly-Lochbaum type transmission line [1] for volume velocity waves. A half sample delay implementation (see [44]) is used so that 16 tube sections, resulting in a vocal tract length of about 17.5 cm, can be used with 16 kHz sampling frequency. Nasal coupling occurs between oral tract tube segments 7 and 8, so that three-tube-junction results into forward propagating wave amplitudes in the 8th oral tract segment and the first nasal tract segment, and into a backward propagating wave amplitude for the 7th oral tract segment. The

reflection coefficient at the glottal end is set to a constant of 0.95. The proper value of the excitation signal is summed to the resulting forward propagating wave at the glottal end. More details on the calculations and the implementation of lip radiation effects and lip protrusion can be found in [20].

The outputs of the nasal and oral tracts are summed and the result is differentiated in order to obtain a pressure signal. The final output pressure signal o of length T is compressed using a hyperbolic tangent function to normalize possible transients in the final output signal, and limit the maximum output value into one:

$$o_{compressed}(t) = \tanh\big(2o(t)\big), \qquad t = 1 \dots T \tag{4}$$

Finally white Gaussian noise is added to the compressed signal so that the resulting signal-to-noise ratio corresponds to 50 dB in the language learning experiments in [19]. This is done in order to diminish the effect of a sound wave of ever decreasing amplitude during closures, when a wave still keeps on propagating in the vocal tract upstream of the closure. Without the noise, clear MFCC features could be extracted during closures, even though the signal power would be inaudible to a human ear, providing an unrealistically strong implication of the closure location.

5 Conclusions

A vocal tract model able to produce a wide variety of different speech sounds, including voiced and unvoiced consonants, fricatives, liquids, vowel sounds and nasals, was created. Phonemes are defined as positions of 9 elementary articulatory parameters, and they are pronounced using related excitation and timing parameters. Properties related to Finnish vowels were investigated and Finnish phoneme system was implemented for Finnish articulatory speech synthesis purposes with pleasing accuracy. The model can also be easily used for babbling utterances without a well-defined phonetic system, making it an ideal tool to investigate infants' speech sound acquisition. The speech rate of the synthesis can be varied, faster speech rate leading in stronger coarticulatory effects. The fundamental frequency was modeled to drift randomly around the user-defined F0-value leading to more natural sounding speech. The vocal tract model has been used in language acquisition experiments in [19].

6 References

[1] J. L. Kelly, C.C. Lochbaum, Speech Synthesis, Proc. 4th Int. Congr. Acoustics, Copenhagen (1962), 1-4.

[2] W.L. Henke, Dynamic Articulatory Model of Speech Production Using Computer Simulation, PhD Thesis, M.I.T. (1966).

[3] P. Mermelstein, Articulatory model for the study of speech production, J. Acoust. Soc. Am. 53(4) (1973), 1070-1082.

[4] C.H. Coker, A model of articulatory dynamics and control, Proc. IEEE 64(5) (1976), 452-460.

[5] S. Maeda, Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model, in: W.J. Hardcastle, A. Marchal (Eds.), Speech production and speech modeling, Kluwer Academic Publishers (1990), 131-149.

[6] E.L. Saltzman, K.G. Munhall, A dynamical approach to gestural patterning in speech production, Ecological Psychology 1 (1989), 333-382.

[7] F.H. Guenther, Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production, Psychological Review 102 (1995), 594–621.

[8] P. Lieberman, E.S. Crelin, & D.H. Klatt, Phonetic Ability and Related Anatomy of the Newborn and Adult Human, Neanderthal Man, and the Chimpanzee, American Anthropologist 74(3) (1972), 287–307.

[9] L.-J. Boë, J.-L. Heim, K. Honda & S. Maeda, The potential Neandertal vowel space was as large as that of modern humans, Journal of Phonetics 30(3) (2002), 465–484.

[10] B. de Boer, Self organization in vowel systems, Journal of Phonetics 28(4) (2000), 441–465.

[11] B.S. Atal, J.J. Chang, M.V. Matthews, J.W. Tukey, Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique, J. Acoust. Soc. Am. 63(5) (1978), 1535-1555.

[12] J.L. Flanagan, K. Ishizaka, K. L. Shipley, Signal models for low bit rate coding of speech, J. Acoust. Soc. Am. 68(3) (1980), 780-791.

[13] T. Toda, A. Black, K. Tokuda, Acoustic-to-articulatory inversion mapping with Gaussian mixture model, Proc. Interspeech (2004), 1129–1132.

[14] S. Hiroya, M. Honda, Estimation of articulatory movements from speech acoustics using an HMM-based speech production model, IEEE Transactions on Speech and Audio Processing 12(2) (2004), 175-185.

[15] S. Ouni, Y. Laprie, Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion, J. Acoust. Soc. Am. 118(1) (2005), 444-460.

[16] H. Rasilo, U. Laine, O. Räsänen, T. Altosaar, Method for speech inversion with large scale statistical evaluation, In Proc. Interspeech'11, Florence, Italy (2011), 2693-2696.

[17] K.L. Markey, The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development, Ph.D. Thesis, University of Colorado, Boulder (1994).

[18] I.S. Howard, P. Messum, Modeling the development of pronunciation in infant speech acquisition, Motor Control 15(1) (2011), 85-117.

[19] H. Rasilo, O. Räsänen, U. Laine, Feedback and imitation by caregiver guides a virtual infant to learnnative phonemes and the skill of speech inversion, submitted for publication, 2013.

[20] H. Rasilo H., Estimation of vocal tract shape trajectory using lossy Kelly-Lochbaum model, Master's thesis, Aalto University, Faculty of Electronics, Communication and Automation (2010).

[21] K. Johnson, P. Ladefoged, M. Lindau, Individual differences in vowel production, J. Acoust. Soc. Am. 94(2) (1993) 701-714.

[22] B.H. Story, I.R. Titze, E.A. Hoffman, Vocal tract area functions from magnetic resonance imaging, J. Acoust. Soc. Am., 100(1) (1996), 537-554.

[23] K. Wiik, Finnish and English Vowels, University of Turku, Doctoral dissertation, 1965.

[24] C. Ericsdotter, Articulatory-acoustic relationships in Swedish vowel sounds, Ph.D. dissertation, Stockholm University, Sweden (2005).

[25] V. Välimäki, Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters. Doctoral thesis. Report no. 37, Helsinki University of Technology, Faculty of Electrical Engineering, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland, (1995).

[26] S. Maeda, The role of the sinus cavities in the production of nasal vowels, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82. 7 (1982) 911-914.

[27] J. Dang J, K. Honda, H. Suzuki, Morphological and acoustical analysis of the nasal and the paranasal cavities, J Acoust Soc Am 96 (1994), 2088–2100.

[28] P. Ladefoged, D.E. Broadbent, Information Conveyed by Vowels, J Acoust Soc Am 29(1) (1957), 98-104.

[29] N. Chomsky, M. Halle, The Sound Pattern of English, Harper and Row, New York, (1968).

[30] B. Lindblom, Spectrographic Study of Vowel Reduction, J. Acoust. Soc. Am. 35(11) (1963), 1773-1781.

[31] S.E.G. Öhman, Coarticulation in VCV utterances: Spectrographic measurements, Journal of the Acoustical Society of America 39 (1966), 151–168.

[32] T. Flash, N. Hogan, The coordination of arm movements: an experimentally confirmed mathematical model. The Journal of Neurosciences 5 (1985), 1688-1703.

[33] T. Cho, P. Ladefoged, Variation and universals in VOT: evidence from 18 languages, Journal of Phonetics 27(2), (1999), 207-229.

[34] D. P. Kuehn & K. Moll, A cineradiographic study of VC and CV articulatory velocities, Journal of Phonetics 4 (1976), 303-320.

[35] K. Suomi, Voicing in English and Finnish stops. A typological comparison with an interlanguage study of the two languages in contact, Publications of the Department of Finnish and General Linguistics of the University of Turku (1980).

[36] I. Maddieson, Phonetic Universals, in the handbook of phonetic sciences (J. Laver & W. J. Hardcastle, editors), Oxford: Blackwells, (1997), 619-639

[37] R. Ogden, Prosodies in Finnish, York Papers in Linguistics 17 (1996), 191-240.

[38] J. Brown & P. Koskinen On Voiced Stops in Finnish, Linguistica uralica 2 (2011), 94-102

[39] A.E. Rosenberg, Effect of glottal pulse shape on the quality of natural vowels, Journal of the Acoustical Society of America 49 (1971), 583–598.

[40] G. Fant, Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations. The Hague: Mouton (1970).

[41] K. N., Stevens, Models for the production and acoustics of stop consonants, Speech Communication 13 (1993), 367-375.

[42] Greenwood, A.R., Goodyear, C.C., Martin, P.A., Measurements of vocal tract shapes using magnetic resonance imaging, Communications, Speech and Vision, IEE Proceedings 139(6) (1992), 553-560.

[43] U. K. Laine, Modelling of Lip Radiation Impedance in z-domain, IEEE Int. Conf. Acoust., Speech, Signal Processing 7 (1982), 1992-1995.

[44] S., Mathur, Variable-length vocal tract modeling for speech synthesis, Master's thesis, Dpt. Of Electrical and Computer Engineering, University of Arizona (2003).

APPENDIX A

Pseudocode for calculating trajectories for the articulatory parameters, based on the input of phonemes and their time instants.

```
Assign following starting values at t=1:

velocities and accelerations of all nine parameters to zero
parameter values at t=1 into desired starting parameter values (e.g. neutral tract)
consonant_end_time to zero
next_target values equal to desired starting parameter values
target_type values to one
all parameters free
prev_target values to one
consonant_approach_starting point values to infinity
next_consonant_known values to zero
next_target_time values to one
current_vowel_target values to desired starting parameter values


for t=2 until end of timeline

      PHASE 1
      if (consonantal gesture ended in t-1) or (release period for any parameter started in t-1)
         set all parameters p free to take new targets
         set next_consonant_known for all parameters p to 0
         set prev_target equal to target_priority for all parameters p
         set target_priority to 1 for all parameters p
      end

   for all parameters p

      PHASE2
      if next_consonant_known(p) equals 0 (the next consonant target in the timeline for p is not
                                                                                          known)
         look for the next consonant target affecting p
         set next_consonant_target_time(p) to the time instant of the next consonant target
         set next_consonant_target(p) to the ATP(p) of the found consonant target
         set consonant_approach_starting_point(p) to max(t, next_consonant_target_time – approach
             duration of the found consonant)
         set next_consonant_known(p) to 1
      end

      PHASE 3
      if t>consonant_approach_starting_point(p) (p should be approaching a consonant target)
         set next_target(p) to ATP(p) of the corresponding consonant
         set next_target_time(p) to the corresponding target time instant of the consonant
         set consonant_end_time equal to the end of the release period of the consonant
         set parameter p busy (to not take new targets before it is freed)
         set target_priority(p) to 2 corresponding to a consonant
      end

      PHASE 4
      if p is free
         if targets are found in range [t+1, t+vowel_lookahead]
            For-loop to go through the found targets from the last towards the first
               If found target is vowel and affects p
                  set current_vowel_target(p) to the corresponding ATP value of the found target
                  set next_target(p) to the corresponding ATP value of the found vowel target
                  set next_target_time(p) to the instant of the found vowel target.
                  Quit the for-loop, so that the last vowel target in the vowel lookahead region is
                                                                                          chosen
               End
            End
         End
```

```
        If any value of prev_target is 2 (i.e. at t-1 a consonant release period started for some
                                                                                    parameter)
           Set next_target_time(p) equal to consonant_end_time
           Set next_target(p) equal to current_vowel_target(p)
           Set parameter p busy
        End

   set time_before_target(p) equal to next_target_time(p)-t
   if time_before_target(p) equals zero (a target has been reached)
      set prev_target(p) equal to one
   end

end
```

Calculate minimum-jerk trajectories from the parameter values at time instant t-1 to the next_target values, using parameters' velocity and acceleration values of time instant t-1 at the starting points and velocity and accuracy of 0 at the goal points. Time_before_target is converted into seconds and used as the time in which the target is intended to be reached.

Assign the obtained position, velocity and acceleration values for all parameters at time instant t.

end