

VocalTractLab 2.3 User Manual

Peter Birkholz

July 10, 2020

Contents

1	Introduction	2
1.1	Purpose	2
1.2	Download, Installation, Requirements	2
1.3	Known issues	3
1.4	How to cite VTL	3
2	Overview	3
2.1	Computational models	3
2.2	Vocal tract model changes	4
2.3	Graphical user interface (GUI)	5
3	Basic signal analysis (Signal page)	6
4	Vocal tract model analysis (Vocal tract page)	8
5	Time-domain simulation of vocal tract acoustics (Time-domain simulation page)	15
6	Gestural score (Gestural score page)	18
6.1	The concept of gestural scores	18
6.2	Editing a gestural score	20
6.3	Copy synthesis	21
6.4	Synthesis from files	22
7	Some typical uses	22
7.1	Analysis of the glottal flow for different supraglottal loads	22
7.2	Comparison of an utterance created by copy-synthesis with its master signal	23
7.3	Create and save a new vocal tract shape	23
7.4	Fitting the vocal tract shape to contours in an image	24
7.5	Change the phonation type for a sentence from modal to breathy	24
7.6	Create a gestural score from a segment sequence	25
A	File formats	25
A.1	Speaker file (*.speaker)	25
A.2	Gestural score file (*.ges)	26
A.3	Segment sequence file (*.seg)	26
B	Major changes from version 2.2 to version 2.3	27
B.1	General changes	27
B.2	Changes in the vocal tract and vocal fold model shapes	28
B.3	Changes in the GUI version	28
B.4	Changes in the API	29

1 Introduction

1.1 Purpose

VocalTractLab (VTL) is an articulatory speech synthesizer and a tool to visualize and explore the mechanism of speech production with regard to articulation, acoustics, and control. It is developed by Dr. Peter Birkholz along with his research on articulatory speech synthesis. With VTL, you can, for example,

- analyze the relationship between articulation and acoustics of the vocal tract;
- synthesize vowels for arbitrary vocal tract shapes with different models for voiced excitation;
- synthesize vowels from an arbitrary set of formants and anti-formants;
- synthesize connected speech utterances based on gestural scores or phone sequences, e.g., for perception experiments;
- analyze the time-varying distribution of pressure and volume velocity within the vocal tract during the synthesis of speech.

However, VTL is not (yet) a text-to-speech system. At the moment, connected utterances can only be synthesized based on gestural scores, as described in Sec. 6. Since version 2.3 it is possible to automatically generate the gestural scores from phone sequences and the corresponding phone durations, which makes it much easier to generate longer utterances.

1.2 Download, Installation, Requirements

Since version 2.3, VTL is distributed as free and open source software under the GNU General Public License (GPL). VTL is written in C++ and developed for the Windows platform. As the code is mostly platform-independent, it should also compile on other platforms with some modifications. To run the distributed binaries on other platforms than Windows, e.g., Linux or Mac, we recommend using a virtual machine, for example VirtualBox (www.virtualbox.org). On Linux and Mac systems, VTL could possibly also run in the tool Wine.

The software is free of charge and available for download as a ZIP file from www.vocaltractlab.de. It needs no special installation. Simply unzip the downloaded file into a folder of your choice. The archive contains one subfolder “GUI” for the full software with a Graphical User Interface, and another subfolder “API” which contains the Application Programming Interface in terms of a dynamic link library.

The GUI version was tested on Windows 10, but can probably run on older Windows versions like Windows 7 and 8, too. A fast computer and a high screen resolution are strongly recommended. Tablet computers and netbooks are generally not suited to work with VTL. On some Windows systems it could be necessary to explicitly install OpenGL. Without OpenGL, the 3D model of the vocal tract will not be displayed properly. The GUI version is started by calling “VocalTractLab2.exe”. The GUI folder contains a couple of other data files, e.g., some dynamic link libraries (*.DLL) and example files. The file “JD2.speaker” is an XML file that defines the default speaker (see Sec. A.1) and is loaded automatically when VTL is started.

With the API you can use a range of VTL function from within your own software. Some Matlab and Python code examples demonstrate how to use it.

Both the GUI and the API folders contain a subfolder “Developer” with the source code. The GUI version contains a project file for Visual Studio 2019, and should be built as “x64 Release”. The only external library that it requires is the cross-platform GUI library wxWidgets 3.1.3 (<https://www.wxwidgets.org/>).

There was no extensive testing of the software, but the parts of the program made available are considered to be relatively stable. Please feel free to report any bugs to peter.birkholz@vocaltractlab.de. Note that VTL comes with no warranty of any kind. The whole software is under continual development and may undergo substantial changes in future versions.

1.3 Know issues

- When the VTL API sources are compiled on Linux systems, there may occur low-amplitude high-frequency bursts in certain regions of the synthesized audio signal. This does not happen with the VTL API on Windows systems, and the reason is not yet clear to us.
- When WAV files are loaded into VTL (GUI version) on Linux in WINE, the audio might sound distorted if its sampling rate is not equal to 44100 Hz.

1.4 How to cite VTL

Currently, the most recent paper that gives a broader overview of the software is

- Birkholz P (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS ONE, 8(4): e60603. doi:10.1371/journal.pone.0060603

If you want to refer to specific models implemented in VTL, please find the appropriate publication in Sec. 2.1.

2 Overview

2.1 Computational models

VTL implements various models, including models for the vocal tract, the vocal folds, the acoustic simulation etc. This section provides references to the most important models, which you should consider reading if you wish to understand them in-depth.

The core of the synthesizer is a 3D articulatory model of the vocal tract that defines the shape of the airway between the glottis and the lips. The model was originally developed by Birkholz (2005) and later refined by Birkholz, Jackèl, and Kröger (2006), Birkholz and Kröger (2006), and Birkholz (2013). The most recent improvements have not been published yet, but some of them are discussed below in Sec. 2.2. The model currently has 19 degrees of freedom (vocal tract parameters) that control the shape and position of the model articulators. Two of these parameters are automatically calculated based on the other parameters, so that there are effectively 17 control parameters now.

For the simulation of acoustics, the (enhanced) area function of the vocal tract is calculated, i.e., the variation of the cross-sectional area along the center line of the model Birkholz (2014). The area function is then transformed into a transmission line model of the vocal tract and can be analyzed in the frequency domain or simulated in the time domain. The basic method for the analysis/synthesis is described by Birkholz and Jackèl (2004) and Birkholz (2005).

For the synthesis of vowels, two approaches are adopted in VTL: One convolves the glottal flow waveform of the Liljencrants-Fant model for the glottal flow (Fant, Liljencrants, and Lin 1985) with the impulse response of the vocal tract transfer function calculated in the frequency domain (similar to the method by Sondhi and Schroeter 1987). The other simulates the acoustic wave motion entirely in the time domain based on a finite-difference scheme in combination with a model of the vocal folds attached to the transmission-line model of the vocal tract. Currently, three vocal fold models are implemented: the geometric model by Birkholz, Drechsel, and Stone (2019), the classical two-mass model by Ishizaka and Flanagan (1972), and a modified two-mass model by Birkholz, Kröger, and Neuschaefer-Rube (2011c) and Birkholz, Kröger, and Neuschaefer-Rube (2011a) with a triangular glottis. These models can be individually parameterized and exchanged in the acoustic simulation. The default and preferred model is currently the geometric model by Birkholz, Drechsel, and Stone 2019. After its publication, it has been extended by an additional “flutter” parameter to generate small pseudo-random f_0 fluctuations according to the proposal by Klatt and Klatt (1990) (page 839). Connected utterances based on gestural scores are always simulated in the time domain, because this method can better account for dynamic and transient

acoustic effects. For the simulation of glottal and supraglottal noise sources, a noise source model based on Birkholz (2014) is used, which has been slightly improved for VTL 2.3.

Coarticulatory effects of the vocal tract are modeled using context-dependent target shapes (vocal tract configurations) for consonants according to Birkholz (2013). This model requires three pre-defined prototype vocal tract target shapes for each consonant: one in each of the contexts /aCa/, /iCi/, and /uCu/. When the consonant is being produced in the simulation, its actual target shape is interpolated between these three prototype targets based on the underlying vowel gesture.

Connected utterances in VTL are defined by gestural scores, a concept taken from articulatory phonology (Browman and Goldstein 1992). A gestural score consists of a number of independent tiers populated with discrete gestures. Each gesture represents the movement of certain articulators (i.e., vocal tract parameters) toward a target configuration. The change of vocal tract parameters in response to these discrete gestures is governed by linear dynamical systems (5th order critically damped low-pass filter). Details of the definition of gestural scores and the mapping to articulatory trajectories is the subject of ongoing research. The basic principles underlying the current implementation are discussed in Birkholz (2007) and Birkholz, Kröger, and Neuschaefer-Rube (2011b).

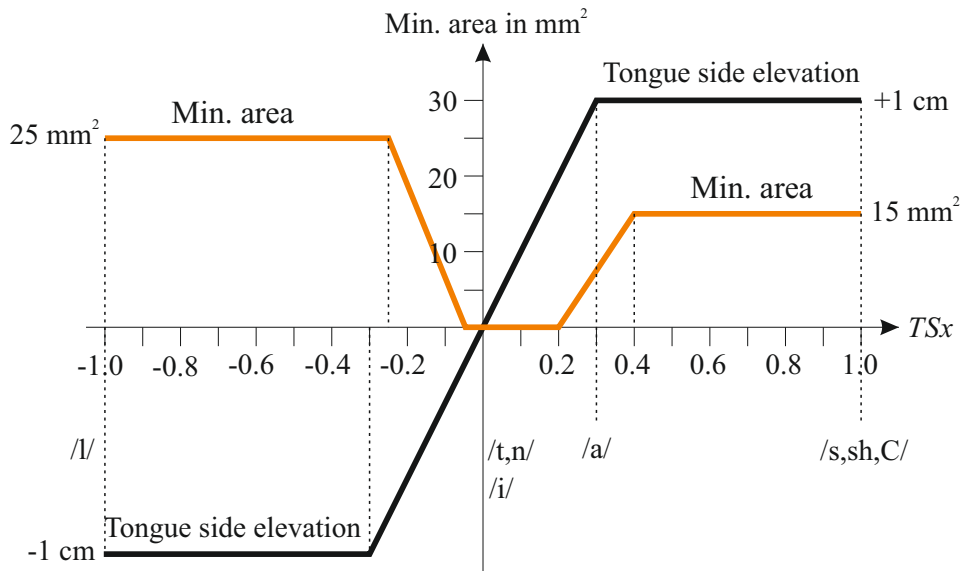


Figure 1: Relation between the tongue side parameters (TS1, TS2, TS3), the elevation of the tongue sides, and the minimal cross-sectional area.

2.2 Vocal tract model changes

For VTL 2.3, the vocal tract model presented by Birkholz (2013) has been further improved. Most importantly, the effective number of control parameters could be reduced to 17 (previously 23). The most important changes concern the parameters TS1, TS2, and TS3. These parameters define the elevation of the tongue sides at three points along the tongue contour: at the tongue root (TS1), at the tongue back and dorsum (TS2), and at the tongue tip and blade (TS3). The value range of TS1 and TS2 is $[0, 1]$, where the tongue is flat (in the coronal plane) for the value 0, and the tongue sides are maximally elevated and stiffened for the value 1. For TS3, the value range is $[-1, 1]$, where values between 0 and 1 cause an increased elevation of the tongue sides as for TS1 and TS2. For $TS3 < 0$ the sides of the tongue tip are lowered to create lateral passages, which are needed for /l/.

The black curve in Fig. 1 shows how the TSx parameters are mapped to an actual elevation of the tongue sides. Here we see that the tongue sides are elevated up to 1 cm for TSx values up to 0.3, and stay at 1 cm for values greater than 0.3. This models a kind of saturation, because the elevation is physiologically constrained to about 1 cm. TSx values greater than 0.3 model an increased stiffening of the tongue muscles (with raised tongue sides) that prevent the formation of a full closure in the vocal tract with the

respective part of the tongue. This leads to a certain minimal cross-sectional area in the vocal tract that is always maintained (orange curve in Fig. 1). For example, for $TS_x > 0.4$ the minimum cross-sectional area in the vocal tract is always 15 mm^2 , which is the typical area of critical constrictions for fricatives. Complete closures (for plosives) can only be made when $-0.05 < TS_x < 0.2$. The minimum areas at the different positions of the tongue are not implemented in the 3D wireframe model of the vocal tract itself, but at the level of the area function. That means that when for example the 3D tongue tip fully penetrates the 3D mesh for the hard palate (as if there was a full closure), but $TS_3 > 0.4$, the area function will have an area of 15 mm^2 (instead of 0) around the position of the tongue tip. This strategy helps to create very precise cross-sectional areas at critical constrictions despite the coarse 3D triangle meshes for the vocal tract walls and articulators.

The left side of the diagram in Fig. 1 only applies to TS_3 (for the tongue tip). For $TS_3 < 0$ the sides of the tongue tip and blade lower down to -1 cm, and at the same time, the minimal cross-sectional area increases up to 25 mm^2 , which is a typical cross-sectional area for lateral passages in laterals according to Narayanan, Alwan, and Haker (1997). This also serves the purpose to ensure a precise cross-sectional area at the position of the tongue tip for laterals.

What does this mean for the vocal tract target shapes for individual speech sounds? For /d,t/ TS_3 must be zero, and for /g,k/ TS_2 must be zero. The vocal tract target shapes for laterals should have $TS_3 = -1$, and the vocal tract shapes for fricatives should have $TS_3 = +1$. For vowel targets, the TS_x parameters should be in the range $[0.0, 0.3]$.

2.3 Graphical user interface (GUI)

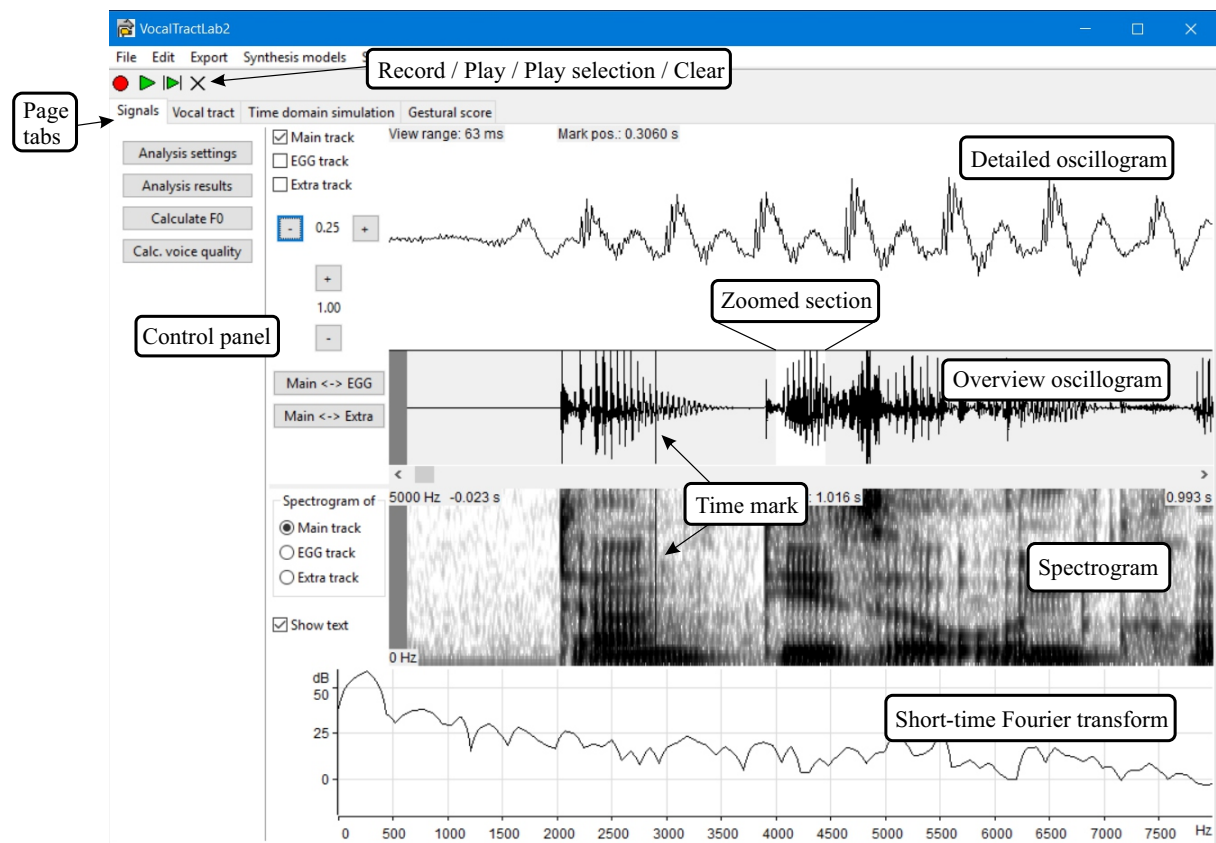


Figure 2: Signal page.

Fig. 2 shows the layout of the graphical user interface of VTL. From top to bottom, it consists of a title bar, a menu bar, a toolbar, and the main window region. The menu bar has the six menus “File”, “Edit”, “Export”, “Synthesis models”, “Synthesis from file”, and “Help”. The items of these menus will be

referred to in the subsequent sections. The toolbar has four tools to record a sound from a microphone, to play the main audio track, to play a selected part of the main audio track, and to clear certain user data, which can be selected from a pull down menu.

The main window region shows one of four pages that can be selected with the page tabs below the toolbar. Each page is dedicated to a certain aspect of the program.

- The *signal page*, which is shown in Fig. 2, is used for the acoustic analysis of speech signals.
- The *vocal tract page* is used for the analysis of the vocal tract model and the synthesis of static speech sounds.
- The *time-domain simulation page* is used for the analysis of vocal tract acoustics and speech synthesis in the time domain.
- The *gestural score page* is used for the creation of gestural scores to synthesize arbitrary utterances and the copy-synthesis of natural utterances.

The following four sections will explain the basic functionality of these pages. In general, each of the pages consists of a main region that contains different displays, and a control panel on the left hand side. In addition to the main window, many functions of VTL are controlled with individual dialogs that can be displayed or hidden on demand. All the dialogs associated with the implemented synthesis models can be called from the menu item “Synthesis models”. The following abbreviations will be subsequently used: LMB = left mouse button; RMB = right mouse button.

3 Basic signal analysis (Signal page)

The signal page provides displays and functions for the basic analysis of audio signals. VTL has three tracks to store digital audio signals. They are called “Main track”, “EGG track”, and “Extra track”. In most cases, all audio analysis and synthesis takes place on the main track. The EGG track is used to store the Electroglottogram (EGG) signal corresponding to a speech signal in the main track. Finally, the extra track can store a third signal for other purposes. For the copy-synthesis of speech based on gestural scores (see Sec. 6), the extra track must contain the original (master) speech signal. Each track represents a buffer of 60 s length with a sampling rate of 44100 Hz and a quantization of 16 bit. The main menu “File” has five items to load and save digital audio signals:

- “Load WAV+EGG (stereo)” loads a stereo WAV file. The left channel is stored in the main track and the right channel is assumed to represent the corresponding EGG signal and is stored in the EGG track.
- “Save WAV+EGG (stereo)” saves the signals in the main track and the EGG track in the selected time range (see below) to a stereo WAV file.
- “Load WAV” loads a mono WAV file to a track of your choice.
- “Save WAV” saves the signal in the selected time range from a track of your choice to a mono WAV file.
- “Save WAV as TXT” saves the signal in the selected time range from a track of your choice as numbers in a text file. This allows the samples of a signal to be easily imported into programs like Matlab or MS Excel.

If the sampling rate of a WAV file to be loaded does not correspond to 44100 Hz, the sampling rate is converted automatically.

There are four displays in the main part of the signal page (cf. Fig. 2): the detailed oscillogram, the overview oscillogram, the spectrogram, and the short-time Fourier transform. Using the checkboxes

next to the oscillogram display you can select the track(s) to display. For each track, the time signal is displayed in a different color. The detailed oscillogram shows a detailed view of the highlighted part in the middle of the overview oscillogram. For one of the tracks, which can be selected next to the spectrogram display, the spectrogram is shown. Both the overview oscillogram and the spectrogram have the same time scale. You can change the scale with the buttons “-” and “+” below the checkboxes next to the detailed oscillogram. The buttons “Main <-> EGG” and “Main <-> Extra” exchange the signals in the corresponding tracks. Use the scrollbar between the oscillogram and the spectrogram to scroll through the tracks. Alternatively, use $\text{Ctrl} + \leftarrow$ or $\text{Ctrl} + \rightarrow$. To select a time range of the signals, right-click in one of the oscillogram displays and select to set the beginning or the end of the range in the context menu. The time range selection is used to define the signal parts to be saved as WAV files or to be played with the “Play selection” button in the toolbar.

When you left-click in the spectrogram or oscillogram displays, you move a time mark (cursor) to the corresponding time in the signal. The short-time Fourier transform (or a different transform, depending on the settings in the Analysis settings dialog in Fig. 3) at the time of the mark is displayed in the bottom part on the page. To change the relative height of the four displays on the page, you can drag the splitter controls right above and below the spectrogram window.

To record a microphone signal, press the red button in the toolbar or press $\text{Ctrl} + \text{R}$. The signal will always be recorded to the main track. To play back the signal in the main track, press the green arrow in the toolbar or press $\text{Ctrl} + \text{P}$. The green arrow between the two vertical lines or the key combination $\text{Ctrl} + \text{[]}$ plays the main track signal in the selected time range. This signal part is played in a loop.

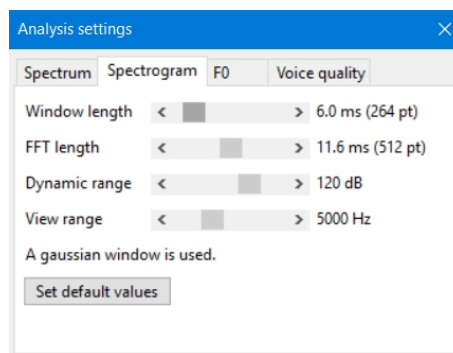


Figure 3: Dialog for analysis settings.

In the control panel of the page are the following buttons:

- “Analysis settings” opens the dialog shown in Fig. 3. This dialog allows setting the parameters used for the calculation of the spectrum in the bottom display, for the spectrogram, and the fundamental frequency (f_0) contour.
- “Analysis results” opens a non-modal dialog where the f_0 at the time of the mark in the oscillogram/spectrogram is displayed in semitones (relative to 1 Hz) and in Hz. Furthermore, it shows measures for the voice quality (phonation type) for the three tracks at the time of the mark.
- “Calculate F0” calculates the f_0 contour in the track of your choice, which can be displayed in the spectrogram (set the corresponding check mark in the analysis settings dialog on the F0-tab).
- “Calc. voice quality” calculates a voice quality contour in the track of your choice similar to the method by Kane and Gobl (2011) (based on the spectral slope), which can be displayed in the spectrogram (set the corresponding check mark in the analysis settings dialog on the voice quality tab).

4 Vocal tract model analysis (Vocal tract page)

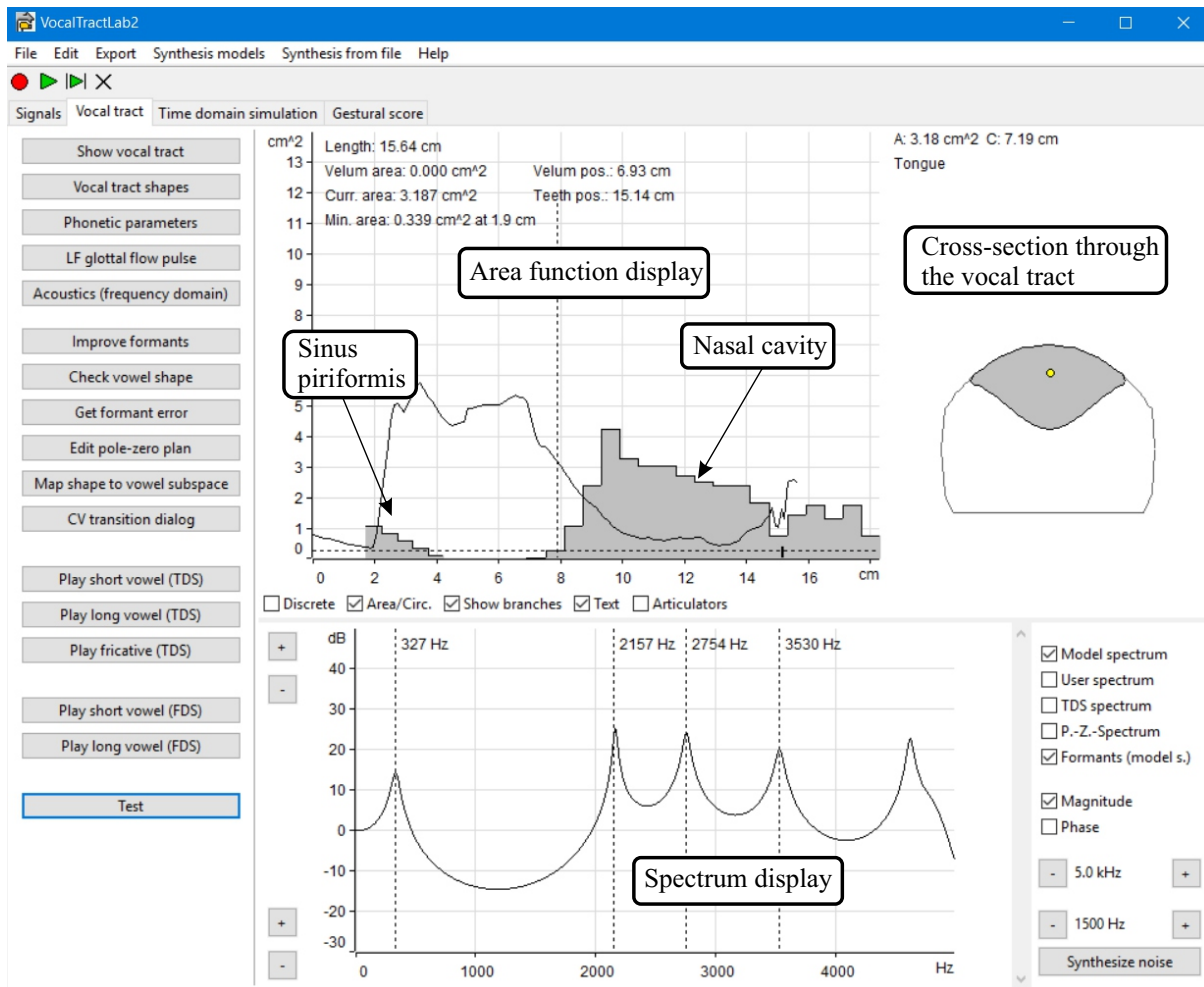


Figure 4: Vocal tract page.

Fig. 4 shows the vocal tract page, which was designed for the articulatory-acoustic analysis of the vocal tract. The actual 3D vocal tract model is shown in a separate dialog (Fig. 5). This dialog is opened automatically when the program is started. If it has been closed, it can be re-opened by selecting the menu item “Synthesis models → Vocal tract model” or pushing the button “Show vocal tract” in the control panel. Some parameters of the model can be changed by dragging the yellow control points with the LMB. A right click on a control point shows the parameter(s) and current value(s) associated with that control point. When no control point is selected, dragging the LMB in the vocal tract display turns the model around two axes. When the mouse is dragged with the RMB, the model can be zoomed in and out.

Three additional vocal tract parameters (TS1, TS2, TS3), which cannot be controlled with the control points, are adjusted using the scrollbars below the vocal tract display. When the checkbox “Automatic TRX, TRY calc.” is checked (default), the vocal tract parameters TRX, TRY, which control the sagittal shape of the tongue root, are automatically calculated based on other parameters. This helps to prevent the creation of unnatural tongue shapes, which could be created otherwise. When you uncheck this checkbox, another control point appears in the vocal tract picture to control TRX and TRY independently from the other parameters.

Display options for the vocal tract model can be found in the bottom part of the dialog. There is also the possibility to attach virtual EMA points (as in Electromagnetic Articulography) to the model articulators. These appear as red points in the vocal tract picture, when “Show EMA points” is checked. The button

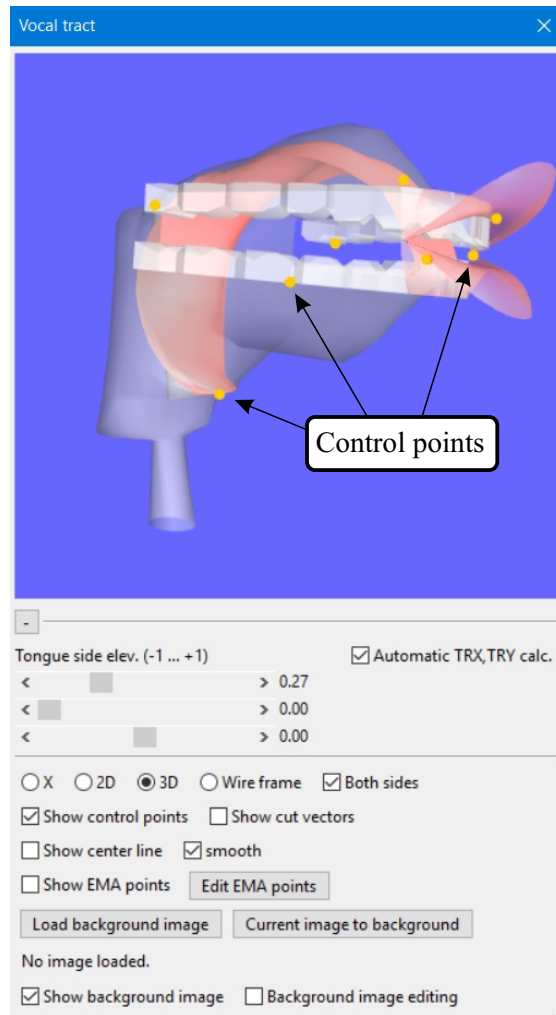


Figure 5: Vocal tract dialog.

“Edit EMA points” opens a dialog, in which the position of the points can be adjusted. The trajectories of these virtual EMA points that would be created by the current gestural score can be exported to a text file with the menu item “Export → EMA trajectories from gestural score”.

Furthermore, you can load an image from a GIF or PNG file with the button “Load background image” to be displayed in the background of the vocal tract model. When you load an image with mid-sagittal contours of a vocal tract, you can try to adapt the articulation of the model to that shown in the image (see Sec. 7.4). When the box “Background image editing” is checked, you can pan and zoom the background image by dragging with the LMB and RMB in the vocal tract display. With the button “Current image to background” the current vocal tract image becomes the background image (with slightly faded colors). This makes it easy to visually compare two vocal tract model shapes, especially if they are displayed as 2D contour images.

The area function corresponding to the current vocal tract shape is shown in the top left display of the vocal tract page. The area functions of the sinus piriformis and the nasal cavity are also included here. Below the display are several options in terms of checkboxes. When the checkbox “Articulators” is checked, the articulators that confine the vocal tract at the anterior/inferior side are shown by means of colors (according to Birkholz (2014)). The information about the articulators plays an important role for the improved acoustic simulation in VTL 2.3. The display right next to the area function shows a cross-section through the 3D vocal tract model, from which the area (grey region) was calculated at a certain position along the center line. The position of this cross-section is marked by the vertical dashed line in the area function and also shown in the vocal tract display, when the box “Show center line” is

checked in the dialog. The position along the center line can be changed by dragging the corresponding control point.

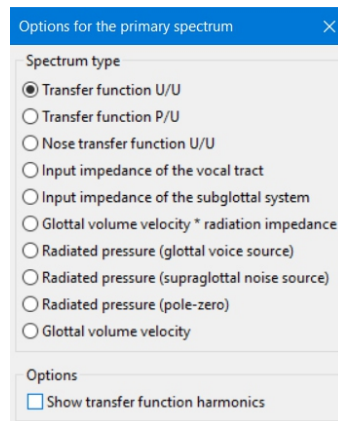


Figure 6: Options for the model spectrum to display.

The display in the bottom part of the vocal tract page shows one or more spectra. The spectrum shown by default is the vocal tract volume velocity transfer function corresponding to the current shape of the vocal tract. This spectrum is calculated in the frequency domain based on the transmission-line model of the vocal tract (cf. Sec. 2.1) with a closed glottis (infinite glottal impedance). The formant frequencies are automatically determined and marked by vertical dashed lines. Note that the formant determination is not reliable when there are zeros (anti-formants) in the transfer function, for example when the velopharyngeal port is open. A left-click in the spectrum display opens the dialog in Fig. 6. Here you can select the kind of model spectrum to display:

- “Transfer function U/U” is the volume velocity transfer function between the (closed) glottis and the lips.
- “Transfer function P/U” is the complex ratio of the radiated sound pressure and the volume velocity at the glottis.
- “Nose transfer function U/U” is the volume velocity at the nostrils divided by the volume velocity entering the nasal cavity at the velo-pharyngeal port.
- “Input impedance of the vocal tract” is the input impedance seen from the glottis.
- “Input impedance of the subglottal system” is the input impedance of the subglottal system seen from the glottis.
- “Glottal volume velocity * radiation impedance” is the product of the line spectrum of the glottal flow of the Liljencrants-Fant model (LF model; see below) and the radiation impedance at the mouth.
- “Radiated pressure (glottal voice source)” is the spectrum of the sound pressure that would be measured in front of the mouth when the vocal tract model is excited by the LF model.
- “Radiated pressure (supraglottal noise source)” is the spectrum of the sound pressure that would be measured in front of the mouth when the vocal tract model is excited by a turbulence noise source that is located at the position of the current cross-section (vertical dashed line in the area function display). The noise source is a dipole (pressure) source shaped with a 2nd-order low-pass filter with a cutoff frequency that can be adjusted in the bottom right corner of the vocal tract page.
- “Radiated pressure (pole-zero)” is the spectrum of the sound pressure that would be measured in front of the mouth when the filter designed in the pole-zero dialog is excited by the LF model.

- “Glottal volume velocity” is the line spectrum of the glottal flow of the Liljencrants-Fant model.

Beside the spectrum selected here (the “model spectrum”), you can display additional spectra by checking the corresponding checkboxes right next to the display. The “user spectrum” is the short-time Fourier transform from the signal page. The “TDS spectrum” is the Fourier transform of the impulse response of the vocal tract calculated using the time-domain simulation. This allows comparing the similarity of the acoustic simulations in the frequency domain and the time domain. The “P-Z-spectrum” is the transfer function defined by a pole-zero plan, that can be created by the user (see below). In the bottom right corner of the vocal tract page is the button “Synthesize noise” (maybe you have to increase the height of the program window to see it), which allows you to synthesize a “fricative” using the current vocal tract shape and a single turbulence noise source at the position that is located at the position of the current cross-section (vertical dashed line in the area function display). The noise source is a dipole (pressure) source shaped with a 2nd-order low-pass filter with a cutoff frequency that can be adjusted with the “+” and “-” buttons right above the “Synthesize noise” button.

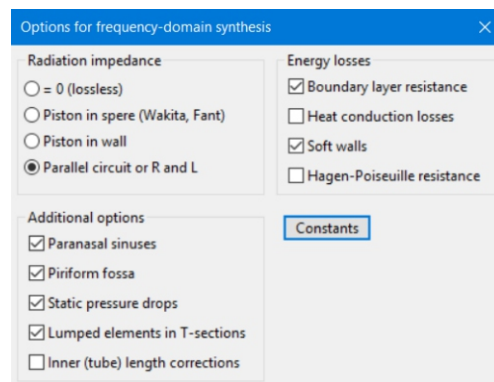


Figure 7: Options for the acoustic simulation in the frequency domain.

The vocal tract transfer function characterizes the acoustic properties of the vocal tract tube between the glottis and the lips. However, there are different options for the calculation of the transfer function from a given area function. The options mainly relate to the loss mechanisms that are considered. They can be changed in the dialog shown in Fig. 7, which is called with the button “Acoustics (frequency domain)” in the control panel. The options are described in detail in Birkholz (2005). If you are not sure what an option means, just use the default value.

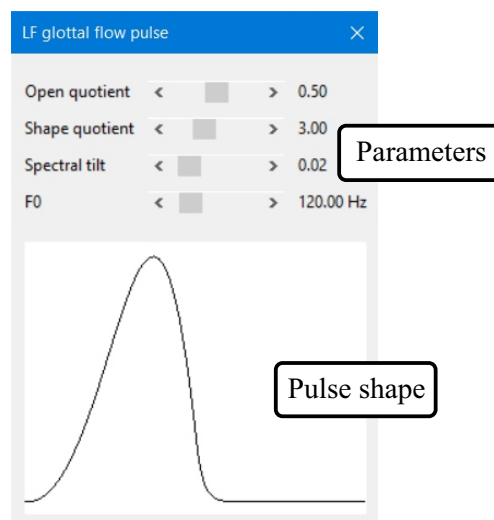


Figure 8: LF glottal flow pulse dialog.

With the buttons “Play short vowel (FDS)” and “Play long vowel (FDS)” in the control panel, the vocal tract model is excited with a sequence of glottal flow pulses to synthesize a short or long vowel in the main track (the previous signal in the main track will be overwritten). The model for the glottal flow pulses is the LF model (Fant, Liljencrants, and Lin 1985). You can change the parameters of the model in the LF glottal flow pulse dialog shown in Fig. 8, which is called with the button “LF glottal flow pulse” in the control panel or from the menu “Synthesis models → LF glottal flow model”. With a left-click in the pulse shape display, you can toggle between the time function of the volume velocity and its derivative with respect to time. The vowels are synthesized by the convolution of the first derivative of the glottal flow pulse sequence with the inverse Fourier transform (i.e., the impulse response) of the vocal tract transfer function. The buttons “Play short vowel (TDS)” and “Play long vowel (TDS)” similarly synthesize and play a short or long vowel using the time-domain simulation together with the geometric glottis model with modal phonation. The button “Play fricative (TDS)” synthesizes and plays a fricative based on the current vocal tract using the “voiceless-fricative” shape of the geometric glottis model.

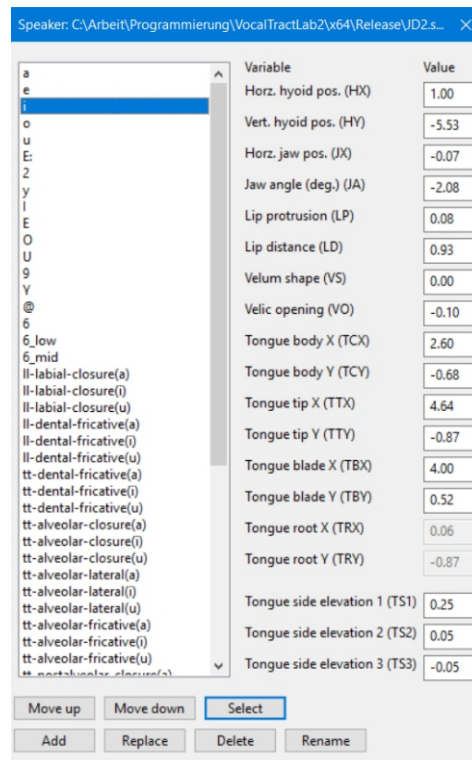




Figure 9: Vocal tract shapes dialog.

The vocal tract model comes with a number of predefined shapes for German vowels and typical consonantal constrictions/closures at different places of articulation. These shapes are used as articulatory targets when gestural scores are transformed into time functions of vocal tract parameters. The shapes are saved in the speaker file (Sec. A.1) and loaded automatically when the program is started. The shapes are managed in the vocal tract shapes dialog shown in Fig. 9, which is called with the button “Vocal tract shapes” in the control panel or from the menu “Synthesis models → Vocal tract shapes”. At the left side of the dialog, the names of the vocal tract shapes are listed. When a shape is selected in the list, the corresponding vocal tract parameters are shown at the right side. How the vocal tract shapes were obtained is described in Birkholz (2013).

The names chosen for the vocal tract shapes of vowels follow the corresponding SAMPA symbols. For the Tiefschwa /ɐ/, there are two variants “6_low” and “6_mid”. The variant to choose for a particular synthesis depends on the phonetic context.

Shapes for consonantal constrictions and closures are sorted by the primary articulator. They start with “ll-” for the lower lip, with “tt-” for the tongue tip, with “tb-” for the tongue body. The following part

of the name indicates the place of articulation (e.g., labial, dental, alveolar), followed by “-closure” for a full oral closure, “-fricative” for a critical constriction, and “-lateral” for a lateral constriction. Each consonantal shape is defined for one of the context vowels /a, i, u/, as indicated by “(a)”, “(i)”, or “(u)” as the final part of the shape name. This naming scheme for consonants is later used in gestural scores to distinguish consonantal gestures with respect to the primary articulator and context. Note that all three context variants for a consonantal configuration must be defined as shapes for the consonant to work as part of gestural scores.

The buttons at the bottom of the dialog allow to (re-)sort the list and to add, replace, delete, rename, and select items. When you click “Add” or “Replace”, the current vocal tract configuration shown in the vocal tract dialog is added to the list or taken to replace an existing item. Click “Delete” or “Rename” to delete or rename a selected item in the list. The button “Select” takes the selected item in the list as the current vocal tract shape. To select a shape from the list and play the corresponding vowel press  in the list of items. To save any changes made to the shape list, you must save the speaker file by pressing  or selecting “File → Save speaker” from the menu.

When you have changed an existing vocal tract shape or added a new one, you might want to check how it behaves in a consonant-vowel transition. To this end, you can open a dialog with the button “CV transition dialog” in the control panel. In this dialog, you can select both a consonant and a vowel in a drop-down box, and then use a scrollbar to change the point on the transition from the consonant to the vowel. Note that the consonant shape here corresponds to the vocal tract configuration of the selected consonant in the context of the selected vowel, i.e., to an interpolated configuration between the consonant prototypes in /a/, /i/, and /u/ context. The vocal tract transfer function and the vocal tract display are changed according to the scrollbar position. When the consonant is a stop consonant, you can use the button “Find release position” in the dialog to find the vocal tract shape along the transition where the oral closure has been released to a specified cross-sectional area.

The button “Improve formants” in the control panel allows the automatic adjustment of the vocal tract shape such that the first two or three formants in the transfer function approach specific values given by the user. A greedy coordinate-descent algorithm tries to find vocal tract parameter values in the vicinity of the current vocal tract shape that change the formant frequencies towards the given values according to Birkholz (2013). This may be interesting, for example, when you want to adapt the formants of a vowel to the realization of that vowel in a different language or dialect. “Max. contour displacement” allows to restrict the shape changes due to the optimization by a certain distance and “Min. cross-sectional area” sets a lower limit for the cross-sectional area along the whole vocal tract that may not be undershot during the optimization. For vowels, the minimal area should not be below 25 mm², because this would cause a critical constriction with a corresponding noise source. Please note that the optimization method is not a full Acoustic-to-Articulatory Inversion, because the vocal tract parameter space is only searched in the vicinity of the current vocal tract configuration.

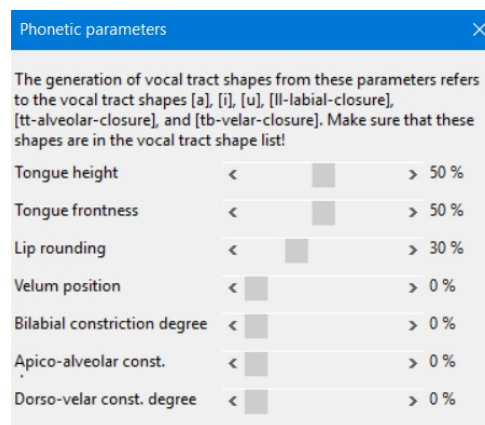


Figure 10: Phonetic parameters dialog.

Besides the level of the elementary vocal tract parameters, which are shown in the vocal tract shapes dialog, VTL provides another higher level of *phonetic* parameters to control the vocal tract shape. This level of parameters is mainly meant for educational experiments. Click the button “Phonetic parameters” in the control panel or select “Synthesis models → Phonetic parameters” from the menu to open the phonetic parameter dialog shown in Fig. 10. There are seven parameters, which can be changed by scrollbars. The parameters “Tongue height” and “Tongue frontness” specify the tongue shape for a vowel. The actual shape is calculated by bilinear interpolation between the predefined shapes for the corner vowels /a/, /i/, and /u/. The degrees of lip rounding and velum lowering are separately specified by the parameters “Lip rounding” and “Velum position”. Finally, the parameters “Bilabial constriction degree”, “Apico-alveolar constriction degree”, and “Dorso-velar constriction degree” can be used to superimpose consonantal constrictions of varying degrees on the vocalic configuration. When a parameter is changed, the corresponding changes in the vocal tract shape, the area function, and the transfer function are immediately displayed.

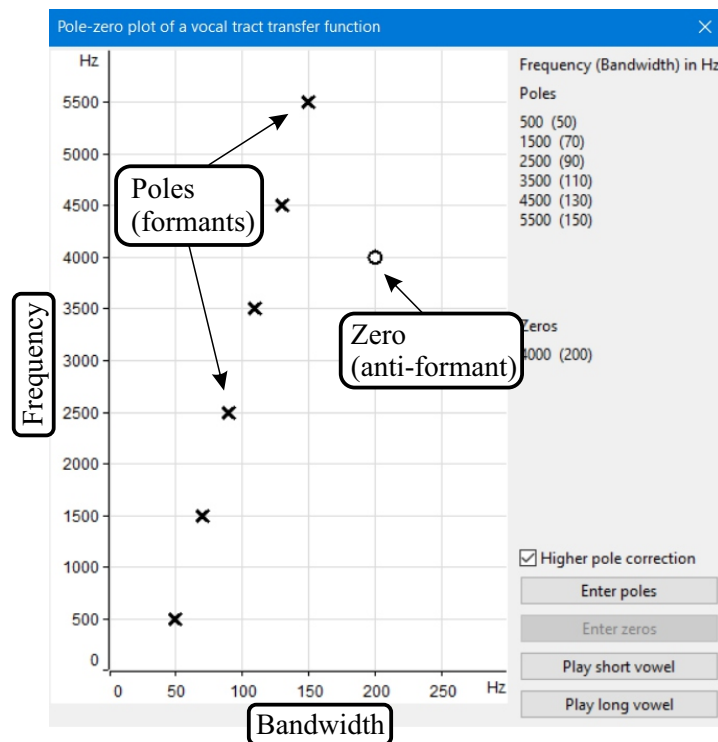


Figure 11: Pole-zero plot.

Independently of the model of the vocal tract, you can specify an arbitrary vocal tract transfer function in terms of a pole-zero plot. This function is mainly meant for educational experiments. Generally, a pole-zero plot displays the poles and zeros of a rational transfer function in the complex plane. In the case of the vocal tract system, the poles correspond to formants and the zeros to anti-formants. Therefore, the pole-zero plot allows you to define a transfer function by means of formant and anti-formant frequencies and bandwidths. To open the pole-zero dialog shown in Fig. 11, click the button “Edit pole-zero plan” in the control panel. The left side shows the location of the poles (crosses) and zeros (circles) in the bandwidth-frequency plane. To change the location of a pole or zero, drag it with the LMB. Right-click in the display to call a context menu to add new or delete existing poles and zeros. When the check box “P-Z-spectrum” next to the spectrum display on the vocal tract page is checked, the transfer function corresponding to the current pole-zero plot is shown. In reality, a vocal tract transfer function has an infinite number of poles. When you want to approximate the effect of the higher poles (the ones above the highest pole that you have placed in the plot) according to Fant (1959) then check the box “Higher pole correction” in the pole-zero dialog (recommended). To play the vowel sound corresponding to the

current pole-zero plot, press the button “Play short vowel” or “Play long vowel”. The audio signal for the vowel is synthesized using the LF glottal pulse model with its current parameter settings and stored in the main audio track. To see the volume velocity transfer function created by the pole-zero pattern, check “P.-Z.-Spectrum” right next to the spectrum picture on the vocal tract page.

5 Time-domain simulation of vocal tract acoustics (Time-domain simulation page)

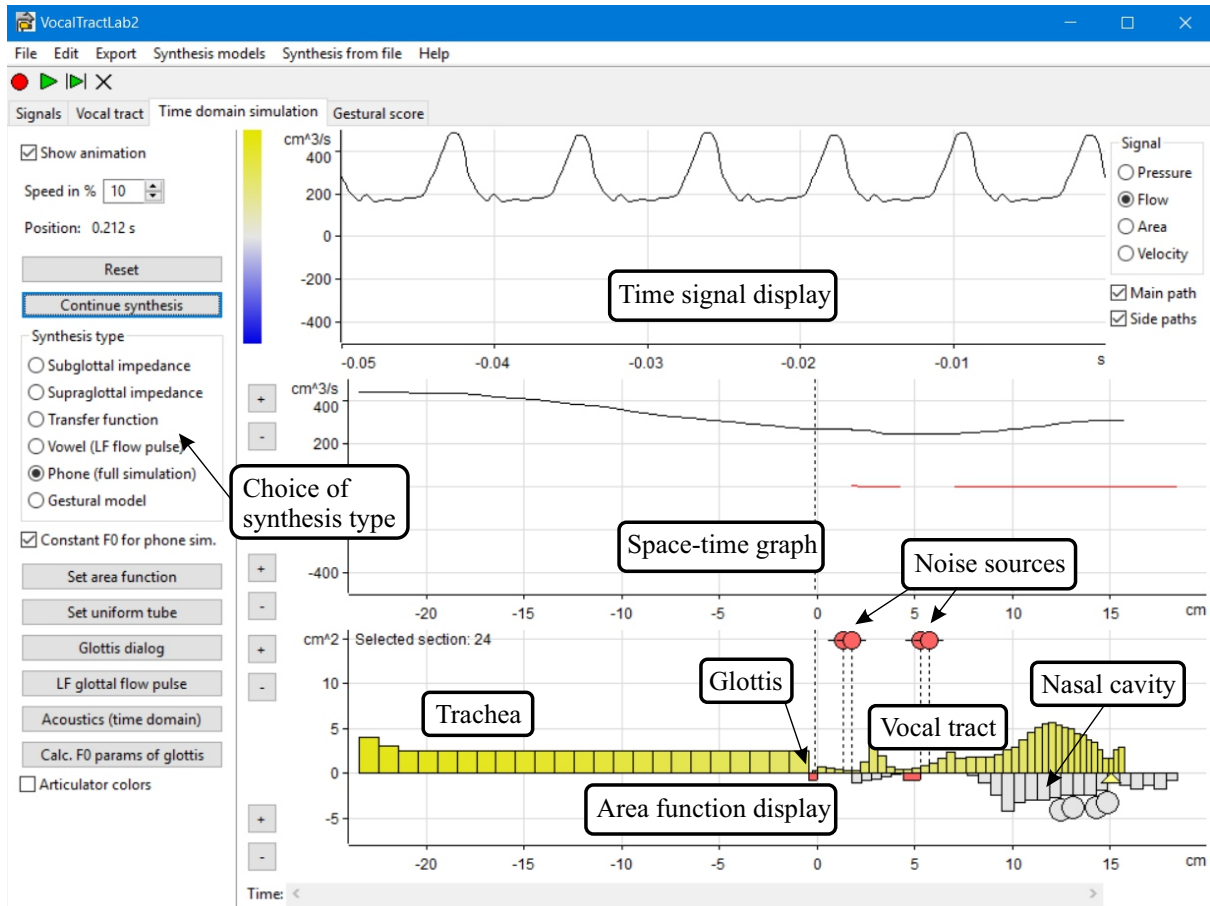


Figure 12: Time-domain simulation page.

Fig. 12 shows the time-domain simulation page, which was designed for the analysis of vocal tract acoustics during the time-domain simulation. Here you can analyze how the pressure and volume velocity (flow) distributions in the vocal tract change over time during speech production.

There are three displays on the page. The area function display at the bottom shows the discrete area function of the vocal system. Each slice of the area function represents a short section of the vocal system tube. The glottis, represented by two very small tube sections, is about in the middle of the display. To the left of the glottis, the trachea is represented by 23 tube sections. The tube sections of the actual vocal tract are shown at the right of the glottis. At the velo-pharyngeal port, the nasal cavity is coupled to the vocal tract tube. Its area function is flipped upside-down. The paranasal sinuses are modeled as Helmholtz resonators according to Dang and Honda (1994) and Dang and Honda (1996) and displayed by four circles. The colors of the tube sections indicate the magnitude of the selected variable at the current time of the simulation. In the top-right corner of the page, you can select the variable to visualize: the sound pressure (“Pressure”), the volume velocity (“Flow”), the cross-sectional area of the tube (“Area”), or the particle velocity of the flow (“Velocity”). In Fig. 12, the flow is visualized.

When there are constrictions in the vocal tract that give rise to turbulence noise during a simulation, this will be indicated by small red circles in the area function display at the position of the generated noise sources. The constrictions are indicated by red horizontal bars under the corresponding tube sections. One of these constrictions is often the glottis that gives rise to a noise source about 1.5 cm downstream from the glottis. A click with the RMB on a constriction (red bar) will provide additional information about the constriction.

Above the area function display is the space-time graph. Here, the selected variable is shown as a curve. The horizontal axis is the position on the vocal tract center line analogous to the area function display. The black curve shows the selected quantity in the trachea, the glottis, and the vocal tract, and the red curve shows it in the nasal cavity and the sinus piriformis. The ordinate of both displays can be scaled with the “+” and “-” buttons left of the displays.

In the area function display, you can select one of the tube sections with the LMB. The selected section is marked with a vertical dashed line. In Fig. 12, the upper glottis section is selected. In the uppermost display of this page, the time signal display, the change of the selected variable in the selected tube section is plotted as a function of time for the last 50 ms. The signal shown in Fig. 12 is therefore the glottal volume velocity in the last 50 ms of the simulation.

The radio buttons in the control panel on the left allow you to choose the “synthesis type” for the simulation:

- *Subglottal impedance* injects a short volume velocity impulse from the glottis into the trachea, records the impulse response of the volume velocity and pressure right below the glottis, and calculates the complex ratio of the Fourier transform of both signals.
- *Supraglottal impedance* injects a short volume velocity impulse from the glottis into the vocal tract, records the impulse response of the volume velocity and pressure right above the glottis, and calculates the complex ratio of the Fourier transform of both signals.
- *Transfer function* injects a short volume velocity impulse from the glottis into the vocal tract and records the impulse response of the volume velocity at the lips. The Fourier transform of this impulse response is the vocal tract transfer function calculated in the time domain. The spectra resulting from the synthesis with this and the previous two options can be shown as a green curve in the spectrum display of the vocal tract page when the checkbox “TDS spectrum” is checked.
- *Vowel (LF flow pulse)* is used for the synthesis of a voiced sound excited by the LF flow pulse model. The button “Set area function” sets the area function for the tube model used for the synthesis to the current area function of the vocal tract model.
- *Phone (full simulation)* is used for the synthesis of a phone with a static vocal tract shape excited by a model of the vocal folds. For this kind of synthesis, you must set the area function with the button “Set area function” and set a rest shape of the vocal fold model used for the synthesis (see below). Depending on the shape of the glottis and the vocal tract, this option can be used to synthesize, for example, static vowels and voiced and voiceless fricatives. When the checkbox “Constant F0 for phone sim.” is checked, the f_0 will be set to a constant value for the whole utterance. Otherwise, the phone will be given a more natural f_0 contour.
- *Gestural model* is used for the synthesis of the utterance defined by the gestural score (Sec. 6). With this option, the area function and the state of the glottis are derived from the gestural score and may change over time. When this option is selected, the scrollbar at the bottom of this page is enabled and allows scrolling through the utterance in time. The corresponding changes in the area function over time are shown in the area function display.

To start the synthesis of the selected type, press the button “Start synthesis / Continue synthesis” in the control panel. When the checkbox “Show animation” is checked, the temporal change of the selected variable can be observed in the displays. The speed of the animation can be changed between 1 and 100%.

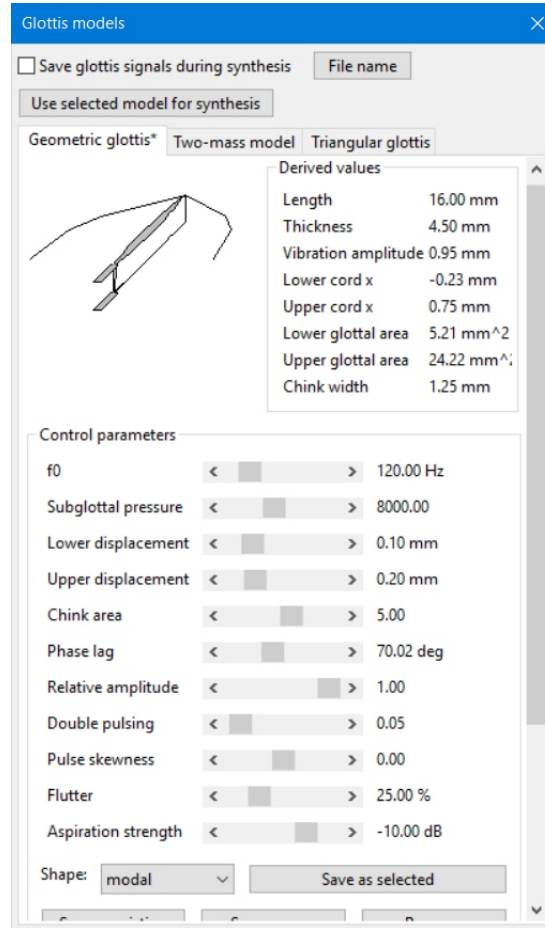


Figure 13: Glottis dialog.

VTL has implemented different models of the vocal folds that can be used for the acoustic synthesis in the time domain. These models are managed in the glottis dialog shown in Fig. 13, which can be called with the button “Glottis dialog” in the control panel or from the menu “Synthesis models → Vocal fold models”. Currently, three vocal fold models are implemented: the geometric model by Birkholz, Drechsel, and Stone (2019), the classical two-mass model by Ishizaka and Flanagan (1972), and the modified (triangular) two-mass model by Birkholz, Kröger, and Neuschaefer-Rube (2011c). Each model is managed on a separate dialog page that can be shown by clicking on the corresponding tab. The model that is currently used for simulations is marked with a star (*) and can be changed, when the button “Use selected model for synthesis” is clicked. Each model has a specific set of *static parameters* and *control parameters*. The static parameters define speaker-specific, shape-independent properties of the model, e.g., the rest length and the rest mass of the vocal folds. On the other hand, the control parameters define the instantaneous shape (posture) and properties of the vocal folds, e.g., the vocal fold tension in terms of the fundamental frequency and the degree of abduction. In the dialog, the control parameters are changed with scrollbars. Individual settings of these parameters can be saved as *shapes*. For example, you can define a shape for modal phonation, a shape for breathy phonation (more abducted vocal folds in the pre-phonatory state), and a shape for voiceless vocal tract excitation (strong abduction). These shapes are referred to in gestural scores where they are associated with glottal gestures (Sec. 6). The current list of shapes for a model can be found in the drop-down list with the label “Shape”. The buttons around the drop-down list allow saving the current setting of control parameters as one of the existing shapes or as a new shape, and to remove shapes from the list. Please note that the control parameters “f0” and “Subglottal pressure”, which exist for all vocal fold models, are not included in the definition of a shape, because these parameters are controlled by independent tiers of gestures in gestural

scores. The parameters of the vocal fold models and the associated shapes are stored in the speaker file “JD2.speaker”, which is loaded automatically when VTL is started. Therefore, to save any changes made to static parameters or shapes, you must save the speaker file by pressing **F2** or selecting “File → Save speaker” from the menu.

The preferred vocal fold model with the most complete set of pre-defined shapes for speech synthesis is the “geometric glottis” model. In contrast to the geometric glottis model, the triangular glottis model is a self-oscillating model. If you should change any of the *static* parameters of this model, this may change the natural fundamental frequency of the model (which is a static parameter by itself). To determine and set the new natural f_0 in this case, press the button “Calc. F0 params of glottis” in the control panel of the page.

When an acoustic simulation of the type “Phone (full simulation)” is started, the selected vocal fold model with the current parameter setting will be used. The time functions of the control parameters and a number of derived parameters of a vocal fold model during an acoustic simulation can be saved to a text file for subsequent analysis. The data are saved to a file when the checkbox “Save glottis signals during synthesis” in the upper part of the dialog is checked. The file name is selected with the button “File name”. The format of the file is as follows: The first line defines the order of the saved parameters, and each following line contains the parameter values of one time step of the simulation (44100 Hz sampling rate). The data can be imported into MS Excel or other programs.

The options for the aero-acoustic simulation in the time-domain can be specified in the dialog that opens when the button “Acoustics (time domain)” in the control panel is pressed. These options are partly similar to those for the frequency-domain simulation, and you should not change them without knowing what they mean.

6 Gestural score (Gestural score page)

The gestural score page allows you to create gestural scores for the synthesis of connected utterances. The page has a control panel at the left side and a signal display and gestural score editor at the right (Fig. 14). The signal display allows the side-by-side comparison of the synthetic speech signal and a natural speech signal and so helps to reproduce natural utterances by articulatory speech synthesis.

6.1 The concept of gestural scores

A gestural score is an organized pattern of articulatory gestures for the realization of an utterance. This concept was originally developed in the framework of articulatory phonology (Browman and Goldstein 1992). While the basic idea is the same in VTL, the specification and execution of gestures differ from articulatory phonology and will be briefly discussed here. In general, a gesture represents movement toward a target configuration of the vocal tract model or the vocal fold model by the participating articulators/parameters. These gestures are organized in eight tiers as shown in Fig. 14. Each tier contains gestures of a certain type. From top to bottom, these are vowel gestures, lip gestures, tongue tip gestures, tongue body gestures, velic gestures, glottal shape gestures, F0 gestures, and lung pressure gestures. Within the same tier, the gestures (grey and white boxes) form a sequence of target-directed movements towards consecutive targets. Some tiers have the exclusive control over a set of vocal tract or vocal fold model parameters, while other parameters are affected by gestures on different tiers.

Glottal shape gestures, F0 gestures, and lung pressure gestures control the parameters of the selected vocal fold model. The lung pressure gestures and the F0 gestures exclusively control the corresponding parameters of the selected vocal fold model. Here, each gesture specifies a target value for the controlled parameter. These target values are sequentially approached. For F0 gestures, the target does not need to be constant in the time interval of a gesture, but may vary as a linear function of time, i.e., it may have a slope. This corresponds to the target approximation model for F0 control by Prom-on, Xu, and Thipakorn (2009). Glottal shape gestures control all remaining control parameters of the vocal fold model. Here, the target associated with a gesture is a shape that was defined for the selected vocal fold model (Sec. 5).

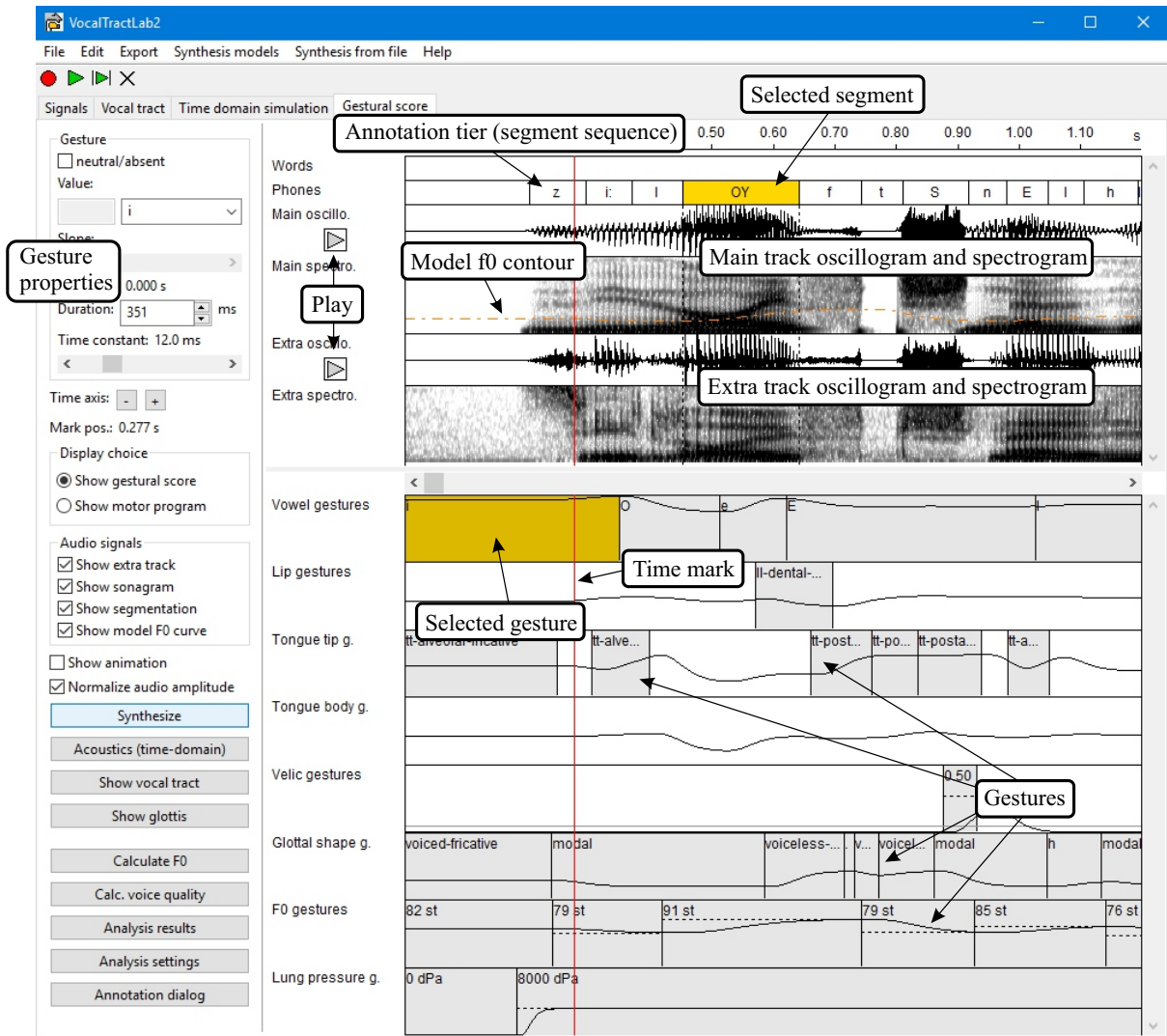


Figure 14: Gestural score page.


In Fig. 14, for example, the score starts with the glottal shape “voiced-fricative” for a slightly abducted vocal fold posture appropriate for a voiced fricative, and then approaches a state with the gesture “modal” to generate modal phonation around the position of the time mark (vertical red line).

Supraglottal articulation is defined by the upper five tiers of gestures. Here, the vowel gestures define basic diphthongal movements of the vocal tract. On these movements are superimposed the constriction forming gestures of the lips, the tongue tip, and the tongue body in the corresponding tiers. In the example for the utterance [zi:lOYftSnElhIn] (in SAMPA notation) in Fig. 14, there are vowel gestures for /i/, /O/, /e/, /E/ and /I/, superimposed with a lip gesture for the labio-dental fricative /f/ and multiple tongue tip gestures for /z/, /l/ etc. Each gesture in the upper four tiers is associated with a particular pre-defined vocal tract shape that serves as the target for the movement. These are the shapes in the vocal tract shapes dialog.



All gestures defined in the gestural score are either “active gestures” or “neutral gestures”. An active gesture is painted as a gray box and specifies a certain target for the vocal tract model or vocal fold model parameters. A neutral gesture is painted as a white box and represents the movement towards a neutral or underlying target. For example, in the tier of lip gestures, the initial neutral gesture is followed by the active gesture for the /f/. In each tier, the time function of a representative vocal tract/vocal fold parameter is drawn. These curves are just meant to illustrate the effect of the gestures on the movements. The curve in the tier of tongue tip gestures shows, for example, the vertical position of the tongue

tip. The horizontal gray line in the bottom part of the tier for velic gestures indicates the dividing line between a closed and a open velo-pharyngeal port. Hence, when the black curve (corresponding to the velum opening/VO parameter of the vocal fold model) is above the gray line, the velo-pharyngeal port is open, and otherwise closed. Please consider that the details of gestural control in VTL are under active development and may undergo major changes in the future.









6.2 Editing a gestural score

Click the LMB on a gesture to select it. The selected gesture is painted in yellow. The properties of the selected gesture can be controlled in the upper part of the control panel. Each gesture has a duration, a time constant, and a value. The duration defines the length of the gesture in the gestural score, and the time constant specifies how quickly the participating articulators reach the target associated with the gesture. A high time constant results in a slow approach, and a low time constant in a fast approach. A value of 12 ms for supraglottal gestures has proven to give satisfactory results in most cases and is the default. The value of a gesture is either a numeric value (the subglottal pressure in dPa for pressure gestures, the f_0 in semitones (rel. 1 Hz) for F0 gestures, or the velum position for velic gestures), or a label. For a glottal shape gesture, the label specifies a pre-defined glottal shape as the target for the gesture. For a vowel, lip, tongue tip, or tongue body gesture, the label specifies a pre-defined vocal tract shape as the target for the gesture. F0 gestures can have a non-zero slope in addition to the target value. The duration of a gesture can also be changed by dragging the vertical border line between two gestures with the LMB. When  is pressed at the same time, only the border is moved. Otherwise, all gestures right from the border are moved along with the border. For velic gestures, F0 gestures, and pressure gestures, the numeric value of the gesture can also be changed by dragging the horizontal dotted line in the gesture box vertically with the LMB.

With a right-click in the gestural score, the red time mark is set to the position of the mouse cursor, the gesture under the mouse cursor is selected, and a context menu is opened. From the context menu, you can choose to insert a gesture or to delete the selected gesture, and you can set a simple “initial” gestural score, or initialize the gestural score from a segment sequence (see Sec. 6.3).

Press  + LMB to set the time mark to the position of the mouse. When the vocal tract dialog or the glottis dialog is shown, the state of the vocal tract or the glottis at the time of the mark is displayed there. You can also drag the mouse from left to right with the LMB while  is pressed. In this case you can observe how the vocal tract shape and the glottis shape change over time.

Additional important keys are:

-  +  simultaneously increases the length of the gestures at the time mark in all tiers by a small amount. In this way you can “stretch” the gestural score at the time of the mark, for example to increase the length of the phone at that time.
-  +  simultaneously decreases the length of the gestures at the time mark in all tiers by a small amount. In this way you can shorten the gestural score at the time of the mark, for example to shorten the length of the phone at that time.
-  +  scrolls the score to the left.
-  +  scrolls the score to the right.

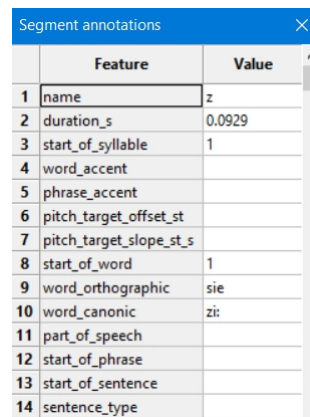
To synthesize the speech signal for a gestural score, click the button “Synthesize” in the control panel (for a fast simulation, uncheck the checkbox “Show animation”). The synthesized audio signal will be stored in the main track and played automatically when the synthesis finished. When the checkbox “Normalize audio amplitude” is checked, the amplitude of the synthesized audio signal will be automatically peak-normalized (to 1 dB below the maximum possible value). Gestural scores can be loaded and saved via the “File menu”.

Please also consider the following hints:

- To generate a glottal stop, you should put a glottal “stop” gesture in the tier with the glottal shape gestures. However, to make the stop sound more natural, you should also rapidly lower (and then rise) the f_0 during the glottal stop. Depending on the f_0 deflection, different degrees of glottalization can be created.
- The precise timing and coordination of the gestures can be really critical for natural-sounding results, especially in consonant clusters. Sometimes shifting a gesture boundary by just 5 or 10 ms can make a big perceptual difference.

When you have loaded or created a gestural score, there are several options to modify the score in a “global” way using the menu items in the menu “Edit”. For example, the menu item “Edit → Change gestural score F0 offset” will change the offset of all pitch targets (gestures) by a certain number of semitones and so generate a higher or lower voice. As another example, the menu item “Edit → Change gestural score duration” allows to stretch or compress the whole score by a given factor in time and so generate faster or slower speech.

6.3 Copy synthesis



	Feature	Value
1	name	z
2	duration_s	0.0929
3	start_of_syllable	1
4	word_accent	
5	phrase_accent	
6	pitch_target_offset_st	
7	pitch_target_slope_st_s	
8	start_of_word	1
9	word_orthographic	sie
10	word_canonic	zi:
11	part_of_speech	
12	start_of_phrase	
13	start_of_sentence	
14	sentence_type	

Figure 15: Segment annotation dialog.

Without a lot of practice, it is quite difficult to create gestural scores for natural-sounding utterances. It is somewhat easier, but still not trivial, to “copy” a natural master utterance with a gestural score, because you can use the acoustic landmarks in the master signal for orientation with respect to the coordination and timing of gestures. Therefore, you can display the synthetic speech signal (on the main track) and the master signal (which must be on the extra track) as oscillograms and spectrograms below each other in the top right panel of this page. In addition, you can show the segment sequence of the master signal above the signals in the annotation tier. You can manually create/edit a segment sequence and load or save it with the menu items “File → Load/Save segment sequence”. You can insert and delete segments by calling the context menu in the annotation tier, and move the borders between segments by dragging them with the LMB (moving all following segments along with the border) or with ⌘ + LMB (to move a border independently). A double-click on a segment opens the segment annotation dialog shown in Fig. 15, where the properties of a segment can be changed. For a description of the properties please refer to Sec. A.3.

Since VTL 2.3 it is also possible to generate a gestural score *automatically* from the segment sequence in the annotation tier. Therefore, call the context menu in the gestural score display and click “Init. score from segment sequence”. The algorithm to map the segment sequence to the score is based on a set of rules that seems to work quite well in most cases, even for consonant clusters. A requirement is that the segment labels are valid SAMPA labels for German. A complete list of possible labels can be found in the source code file Sampa.cpp. The first segment of a sequence should always be a pause of about 100 ms (the label for the pause is an empty string), because the gestures for the first non-pause segment usually

start earlier than the segment. The generation of gestural scores for other languages than German should also be possible with this method when you first replace non-German SAMPA labels in the annotation tier with German labels, start the transformation into a gestural score, and finally substitute the “wrong” gestures with the needed gestures for the target language. The F0 gestures are not affected when you transform a gesture sequence into a gestural score.

Which parts of the display in the top right part of the gestural score page are shown can be controlled in the button group “Audio signals” in the control panel. The checkbox “Show model F0 curve” allows showing the model f_0 curve as a red dashed line in the spectrogram of the main track to compare it with the measured f_0 contour of the master signal in the extra track.

6.4 Synthesis from files


Besides the use of gestural scores, there are two other ways to generate connected utterances on the basis of the time-domain simulation:



1. Generate the speech signal from a sequence (i.e. concatenation) of vocal tract tube states with the menu item “Synthesis from file → Tube sequence file to audio”. The vocal tract tube states (i.e. enhanced area functions) must be defined in a TXT file with a rate of about 400 states/s. You can create such a tube sequence file from a gestural score with the menu item “Synthesis from file → Ges. score to tube sequence file”. The file format is explained in the header lines of such a file. In principle, this kind of synthesis can be used to generate speech from any source of vocal tract area functions, e.g. area functions measured from MRI data of the vocal tract.
2. Generate the speech signal from a sequence (i.e. concatenation) of vocal tract model and glottis model states with the menu item “Synthesis from file → Tract sequence file to audio”. The states of the vocal tract and glottis models must be defined in terms of control parameter values in a TXT file with a rate of about 400 states/s. You can create such a file from a gestural score with the menu item “Synthesis from file → Ges. score to tract sequence file”. The file format is explained in the header lines of such a file. In principle, this kind of synthesis can be used to generate speech from arbitrary time-functions of control parameters of the glottis and vocal tract models, bypassing the gestural score model.

Both synthesis methods described above can also be used to make and analyze specific manipulations to area functions or control parameters of the vocal tract and glottis models, which are not possible with the gestural scores. For example, you can generate a gestural score for a certain word, then export the control parameters of the vocal tract and glottis models as a tract sequence file, manipulate one or more time functions of the control parameters in these files outside of VTL, and then use the menu item “Synthesis from file → Tract sequence file to audio” to generate audio from the manipulated TXT file.

7 Some typical uses

7.1 Analysis of the glottal flow for different supraglottal loads

1. Select the time-domain simulation page with the corresponding tab below the toolbar.
2. Select one of the two tube sections of the glottis between the tracheal sections and the vocal tract sections with a left-click in the area function display. The vertical dashed line that marks the selected section should be placed as in the area function display in Fig. 12. The time signal of the selected section and the selected variable will be shown in the upper display during simulations.
3. Select the radio button “Flow” in the top right corner of the page.
4. Open the vocal tract shapes dialog with the menu “Synthesis models → Vocal tract shapes”. Double-click on the shape “a” in the list to make it the current vocal tract configuration and close the dialog with .


5. Click the button “Set area function” in the control panel to set the area function of the vocal tract for the acoustic simulation.
6. Select the radio button “Phone (full simulation)” in the control panel, check the box “Show animation”, and press “Start synthesis”. You should now see the time function of the glottal flow in the upper display of the page. Wait for the simulation to finish to hear the vowel. After the simulation, you can replay the vowel (which is stored in the main track) with  + .
7. Select the vowel “u” from the list of vocal tract shapes, press “Set area function”, and start the synthesis again. Observe how the glottal pulse shape differs from that for the vowel “a” just because of the different supraglottal configuration.
8. For a detailed analysis of the glottal flow waveforms you should write it to a file by checking the checkbox “Save glottis signals during synthesis” in the dialog with the glottis models, and select a file for the data. When you now synthesize again, the glottis data are written to the file.

7.2 Comparison of an utterance created by copy-synthesis with its master signal

1. Select the gestural score page with the corresponding tab below the toolbar.
2. Load the gestural score file “example01.ges” with the menu item “File → Load gestural score”. The gestural score appears in the right bottom part of the page.
3. Press the button “Synthesize” in the control panel and wait for the synthesis to finish. The oscillogram and spectrogram of the synthesized signal appear in the top right display.
4. Load the audio file “example01-orig.wav” to the *extra track* with the menu item “File → Load WAV”.
5. Load the segment sequence “example01.seg” with the menu item “File → Load segment sequence”.
6. Check the checkboxes “Show extra track” and “Show segmentation” in the control panel to show the segment sequence and the master audio signal in the main track.
7. Press the play buttons next to the oscillograms of the original and the synthetic signal to compare them perceptually.
8. Compare the spectrograms of the original and the synthetic signal. You can change the height of the spectrograms by dragging the splitter control right above the scrollbar for the time, which separates the upper and lower displays. Furthermore, zoom in or out with the buttons “+” or “-” in the control panel.
9. Right-click on a segment in the annotation tier to open the context menu and select “Play segment (main track)” or “Play segment (extra track)” to play the corresponding audio signal parts.

7.3 Create and save a new vocal tract shape

1. Select the vocal tract page with the corresponding tab below the toolbar.
2. Open the vocal tract shapes dialog and the vocal tract dialog with the buttons “Vocal tract shapes” and “Show vocal tract” in the control panel.
3. Select a shape from the list, for example “a”, with a double-click on the item.

4. Drag around some of the yellow control points in the vocal tract dialog. This changes the corresponding vocal tract parameters. Observe the changes in the area function and the vocal tract transfer function. Press the button “Play long vowel (TDS)” in the control panel to hear the sound corresponding to the current articulation.
5. Press the button “Add” in the vocal tract shapes dialog to save the current vocal tract configuration to the list of shapes. Type “test” as the name for the new shape and click “OK”.
6. Press  to save the speaker file “JD2.speaker” to permanently save the new shape. When you now close VTL and start it again, the speaker file containing the extended list of shapes is automatically loaded.

7.4 Fitting the vocal tract shape to contours in an image

1. Select the vocal tract page with the corresponding tab below the toolbar.
2. Click the button “Show vocal tract” in the control panel to show the vocal tract dialog.
3. Click the button “Load background image” in the dialog and load the file “vowel-a-outline.gif”. This image shows the outline of the vocal tract for the vowel /a/ obtained from MRI data. The image can be seen behind the 3D vocal tract model.
4. Click the radio button “2D” in the vocal tract dialog to show the mid-sagittal outline of the vocal tract model.
5. Check the checkbox “Background image editing” at the bottom of the dialog. Now drag the background image with the LMB and scale it by dragging the RMB such that the outline of the hard palate and the rear pharyngeal wall coincide in the background image and the model. Then uncheck the checkbox “Background image editing” again.
6. Now try to drag the control points of the model so that the shape of the tongue, the lips and so on correspond to the background image. Press the button “Play long vowel (TDS)” in the control panel of the vocal tract page to hear the corresponding vowel. When the contours coincide well, you should hear an /a/.
7. Click the button “Vocal tract shapes” to open the vocal tract shapes dialog and double-click on the shape “a-raw”. The parameters of this shape were manually adjusted to coincide as accurately as possible with the background image.

7.5 Change the phonation type for a sentence from modal to breathy

1. Select the gestural score page with the corresponding tab below the toolbar.
2. Load the example gestural score “example01.ges” with the menu item “File → Load gestural score”.
3. Click the button “Synthesize” to synthesize the sentence with modal phonation.
4. Click the menu item “Edit → Substitute gestural score glottal shapes” and enter “modal” in the first, and “breathy” in the second dialog that appear. Then all “modal” glottal shape gestures are substituted by “breathy” glottis shapes.
5. Click the button “Synthesize” to synthesize the sentence again. Now the sentence should sound breathy.

7.6 Create a gestural score from a segment sequence

1. Select the gestural score page with the corresponding tab below the toolbar.
2. Load the example segment sequence file “example05.seg” with the menu item “File → Load segment sequence”.
3. Right-click in the lower display with the gestural score to open the context menu, and select “Init. score from segment sequence”. This creates the gestures needed to generate a speech signal with the intended phones and phone durations.
4. Click the button “Synthesize” to synthesize the gestural score. Apart from a small accent at the beginning of the sentence, it generates a flat intonation.
5. Insert additional F0 gestures to generate a more natural f_0 contour, and check the generated intonation by synthesizing the audio signal again.

A File formats

A.1 Speaker file (*.speaker)

The speaker file is an XML file that defines a model speaker. The definition comprises the anatomy of the speaker, the vocal tract shapes used to produce individual phones, as well as model properties for the glottal excitation. The default speaker file is “JD2.speaker”. It must be located in the same folder as “VocalTractLab2.exe” and is loaded automatically when the program is started.

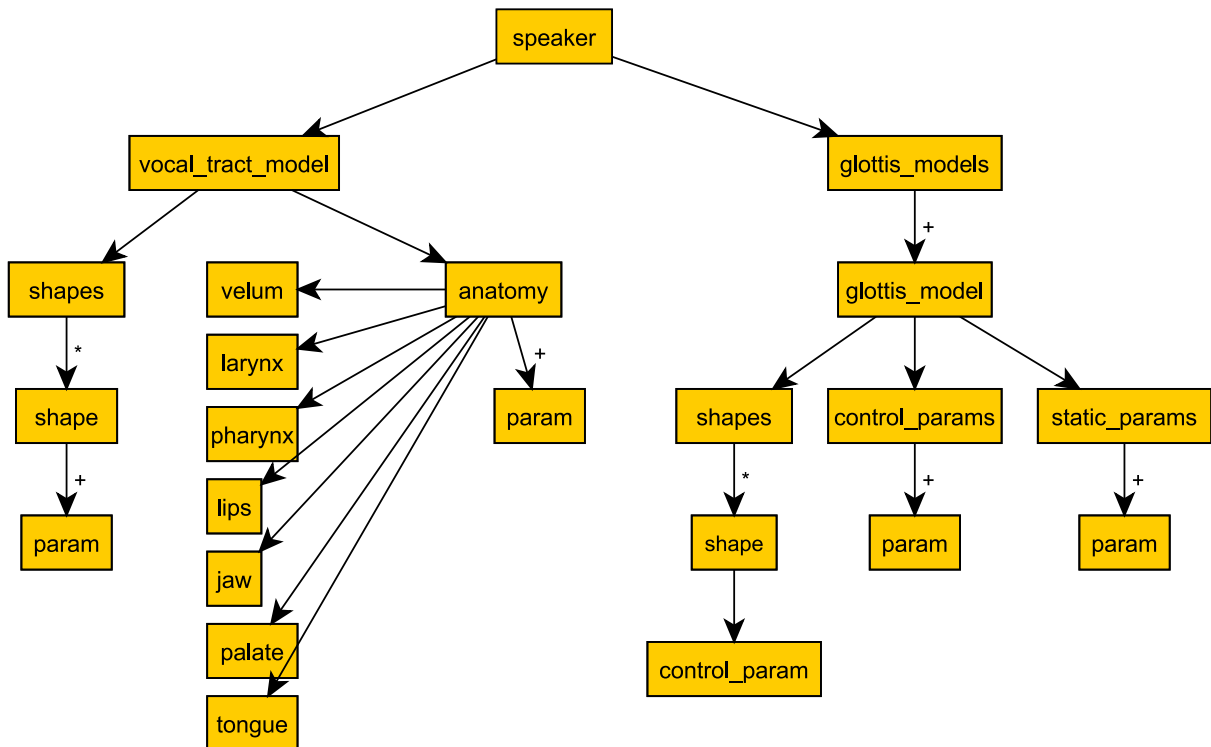


Figure 16: DTD tree of the speaker file.

The document type definition (DTD) tree in Fig. 16 shows the structure of the speaker file. The root element `speaker` has two child elements: `vocal_tract_model` and `glottis_models`. The `vocal_tract_model` element defines the supraglottal part of the vocal tract. It has the child elements `anatomy` and `shapes`. The `anatomy` element defines the shape of the rigid parts of the vocal tract (e.g.,

jaw and palate), properties of the deformable structures (e.g., velum and lips), and the list of parameters that control the articulation. The `shapes` element contains a list of `shape` elements. Each `shape` defines a vocal tract target configuration for a phone in terms of vocal tract parameter values. These are the shapes that are used when gestural scores are transformed into actual trajectories of vocal tract parameters. The element `glottis_models` defines the properties of one or more vocal fold models, which can be used interchangeably to generate the glottal excitation of the vocal tract model in time-domain simulations of the acoustics. Each of the vocal fold models is defined by an element `glottis_model`, which in turn contains a list of glottal shapes (`shapes`), control parameters (`control_params`), and static parameters (`static_params`). The static parameters define the speaker-specific properties of the model, i.e., the parameters that don't change over time (e.g., the rest length and the rest mass of the vocal folds). They are analogous to the anatomic part of the supraglottal vocal tract. The control parameters define the properties that are controlled during articulation, e.g., the vocal fold tension or the degree of abduction. The `shapes` element contains a list of `shape` elements, each of which defines a setting of the control parameters, e.g., a setting for modal phonation and a setting for voiceless excitation (abducted vocal folds). These shapes are analogous to the vocal tract shapes for supraglottal articulation.

A.2 Gestural score file (*.ges)

A gestural score file is an XML file that defines a gestural score. The document type definition (DTD) tree in Fig. 17 shows the structure of the file. The root element is `gestural_score`. There are eight tiers of gestures in a gestural score, each of which is represented by one `gesture_sequence` element. Each gesture sequence comprises a set of successive gestures of the same type (e.g., vowel gestures or velic gestures). The start time of a gesture is implicitly given by the sum of durations of the previous gestures of the sequence.

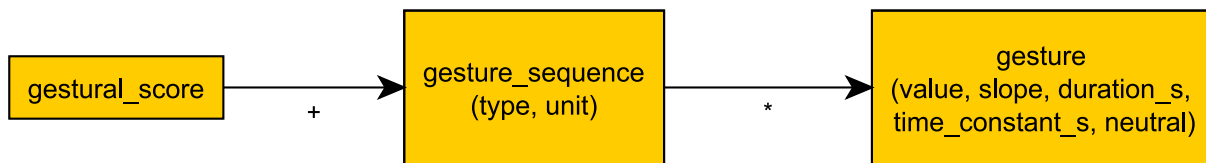


Figure 17: DTD tree of the gestural score file.

A.3 Segment sequence file (*.seg)

A segment sequence file is a text file used to represent the structure and metadata of a spoken utterance in terms of segments (phones), syllables, words, phrases and sentences. The following example illustrates the structure of the file for the word “Banane” [banan@]:

```

duration_s = 0.137719;
name = !b; duration_s = 0.013566; start_of_syllable = 1; start_of_word = 1;
    word_orthographic = Banane; word_canonic = banan@;
name = a; duration_s = 0.072357;
name = n; duration_s = 0.114149; start_of_syllable = 1;
name = a; duration_s = 0.212388;
name = n; duration_s = 0.068383; start_of_syllable = 1;
name = @; duration_s = 0.195274;
  
```

The first text line defines the length of the utterance in seconds. Each following line defines one segment in terms of its name (e.g., SAMPA symbol) and duration. When a segment is the first segment of a syllable, a word, a phrase, or a sentence, this can be indicated by additional attributes in the same text line where the segment is defined. For example, `start_of_word = 1` means that the current segment is the first segment of a new word. The following list shows all possible attributes for a segment:

- `name` defines the name of the segment.

- `duration_s` defines the duration of the segment in seconds.
- `start_of_syllable` set to 1 marks the start of a new syllable.
- `word_accent` set to 1 places the stress in the word on the current syllable. Other numbers can be used to indicate different levels of stress.
- `phrase_accent` set to 1 places the stress on the current word within the phrase.
- `start_of_word` set to 1 marks the start of a new word.
- `word_orthographic` defines the orthographic form of a word.
- `word_canonic` defines the canonical transcription of a word.
- `part_of_speech` defines the part of speech of a word. Any values can be used here, e.g., verb, noun, or adjective.
- `start_of_phrase` set to 1 indicates the start of a new phrase.
- `start_of_sentence` set to 1 indicates the start of a new sentence.
- `sentence_type` can be used to indicate the type of sentence, e.g., it can be set to question or statement.

Apart from `name` and `duration_s`, the use of the attributes is optional.

B Major changes from version 2.2 to version 2.3

Due to the many changes since version 2.2 of VTL, neither the gestural score files nor the speaker file are compatible with the previous version.

B.1 General changes

- VTL is now open source under the GNU GPL license. There is no need to register anymore for the full range of functions.
- The synthesized audio signals now have a sampling rate of 44100 Hz (previously 22050 Hz).
- The synthesized audio signals are now bandlimited to 12 kHz (previously 8 kHz). Although the plane-wave simulation of acoustics is not physically correct above about 5 kHz, the "wideband" synthetic audio sounds still better than for a more limited bandwidth.
- A new “geometric glottis” model has been added and replaces the previous “Titze model”. The new model is the default model for the synthesizer and proved to provide a better synthesis quality than the self-oscillating “triangular glottis” model (which is still available, as before).
- The geometric glottis model has a new “flutter” parameter to control the degree of pseudo-random f_0 fluctuations.
- The constants for the mass, resistance and compliance of the vocal tract walls have been changed so that the first formant is less damped and that we get sufficiently long voicing during the closure of voiced plosives.
- All noise sources in the vocal tract are dipole sources with a spectral envelope that corresponds to a critically-damped 2nd-order low-pass filter. The cutoff-frequency and the gain of a source depend on the place of articulation, on the flow conditions in the constriction, and on whether a plosive or fricative is produced (the noise source spectrum was found to be different for plosives and fricatives at the same place of articulation).

- The (previously disabled) vocal tract parameter “WC” has been removed.
- The minimum-area parameters MA1, MA2 and MA3 have been removed (they were redundant).
- There are now only three “tongue side” parameters (TS1, TS2, and TS3) instead of four, which is sufficient to model the elevation of the tongue sides. TS2 and TS3 now also control the minimum area at the tongue blade, tongue tip, and tongue dorsum regions ($TS2/3 > 0$) and potential laterality ($TS3 < 0$).
- The sampling rate of the parameter time functions in the gestural scores has been increased to 400 Hz (time step of 2.5 ms) to allow a more precise timing of articulatory events.
- To consider intrinsic velocity differences of the articulators, corresponding direction-dependent velocity factors have been added to the anatomic part of the speaker file. In the `<param>` element, there are now the two additional attributes “positive_velocity_factor” and “negative_velocity_factor”, which are usually 1.0. They are used to modify the time constants of the target sequences (“motor program”) generated from the gestural score. Do not change them when you do not exactly know what they mean!

B.2 Changes in the vocal tract and vocal fold model shapes

- The vocal tract and glottis shapes have been partially renamed for better consistency.
- The glottal shapes have been carefully adjusted for good synthesis results and realistic glottal areas.
- The lateral /l/ now has a virtual tongue tip target.
- /m,n,ŋ/ and all plosives also use virtual targets for the primary articulator so that the closure (release) happens with a high velocity in the middle part of the transition.
- The glottal stop shape has a relative amplitude smaller than zero.
- There are two new vocal tract shapes “6_low” and “6_mid” for vocalized /r/ allophones in different contexts.
- The synthesis of the primary German diphthongs works well using the German monophthongs (/aU/ = /a:/+/o:/; /aI/ = /a:/+/e:/; /OY/ = /O:/+/e:/). Therefore, the previous individual vocal tract shapes for diphthongs have been removed.
- The size of the piriform fossae has been changed to a total volume of 1.5 cm³ and a length (including acoustic end correction) of 2.5 cm. This increases the frequency of the major antiresonance in the vocal tract transfer function to 5.2 kHz, which makes the synthesis sound less bass-like.
- There are now also vocal tract shapes for a *postalveolar* closure and lateral. These shapes are used in the automatic generation of gestural scores (from segment sequences) in the neighbourhood of /ʃ,ʒ/ to account for realistic coarticulation.

B.3 Changes in the GUI version

- The time constants of the gestures in the gestural scores are now by default 12 ms (previously 15 ms), which is more appropriate for a normal speaking rate.
- On the vocal tract page next to the spectrum is a new button to synthesize a fricative with a single dipole noise source at the position specified by the cut plane mark in the area function (frequency-domain synthesis). The cutoff frequency of the shaping filter (2nd-order low-pass) of this noise source can be adjusted with the buttons right above it. The spectral envelope of the radiated noise due to this source can also be shown in the spectrum display using the context menu.

- On the time-domain-simulation page, a right-click on a constriction (red horizontal bar) in the area function during the synthesis will show data about the constriction, the flow etc.
- On the gestural score page you can drag the boundary between two segments in the segment (annotation) tier without moving the other boundaries when you press SHIFT while dragging.
- It is now possible to automatically translate a given sequence of phones (SAMPA labels) and phone durations given in the annotation tier into a gestural score that produces a corresponding audio signal using the context menu of the gestural score editor (major novelty!).
- The upper frequency limit of the spectrograms on the gestural score page is now adjustable up to 12 kHz with the context menu.
- On the gestural score page you can play the part of the main or extra track in the temporal domain of the selected segment in the annotation tier with the context menu.
- There is a new menu item “Synthesis from file” in the main menu, which allows you to perform a (blocking) synthesis of speech from a gestural score file, from a tube sequence file, or from a tract sequence file. The latter two file types are text files that can be created from gestural scores in the same menu. They are meant for external manipulations of vocal tract or tube parameters of utterances.
- In the vocal tract model dialog, there is now a button to make the current vocal tract the background image. In this way you can more easily compare two vocal tract shapes visually.

B.4 Changes in the API

- Some of the functions have been renamed for better consistency.
- There is a new function `vtlTractToTube()` to obtain the area function (and additional information) for a given vector of vocal tract parameters.
- The API functions can be “muted” so that they do not write any output to the console, except in case of errors.
- Calling the synthesis functions with exactly identical parameters will now lead to exactly the same synthetic audio signals. Previously, different starting conditions of the random number generator led to different realizations of the turbulence noise.
- There is a new function `vtlExportTractSvg()` to export the midsagittal vocal tract contour for a given vocal tract parameter vector as an SVG file (scalable vector graphics).
- The function `vtlGesturalScoreToAudio()` now requires that the API has been initialized before (as for all other functions).
- There is a new function `vtlSegmentSequenceToGesturalScore()` that translates a given sequence of phones (SAMPA labels) and phone durations into a gestural score that produces a corresponding audio signal. The pitch tier of the gestural score is initialized with a flat intonation here and must be manipulated later for a realistic intonation.

Acknowledgments

I thank Paul Krug and Rémi Blandin for proofreading this manuscript. Some parts of the software were contributed by Thomas Uhle, who helped to make the code ready to compile also on Linux systems, and Johann Marwitz, who developed a faster solver for the linear system of equations in the time-domain simulation of the acoustics. Parts of the research leading to the development of VocalTractLab were funded by the German Research Foundation (DFG), grants JA 1476/1-1 and BI 1639/4-1.

References

- Birkholz, P. (2013). “Modeling consonant-vowel coarticulation for articulatory speech synthesis”. In: *PLOS ONE* 8.4, e60603.
- Birkholz, Peter (2005). *3D-Artikulatorische Sprachsynthese*. Logos Verlag Berlin.
- (2007). “Control of an Articulatory Speech Synthesizer based on Dynamic Approximation of Spatial Articulatory Targets”. In: *Interspeech 2007 - Eurospeech*. Antwerp, Belgium, pp. 2865–2868.
- (2014). “Enhanced area functions for noise source modeling in the vocal tract”. In: *Proc. of the 10th International Seminar on Speech Production (ISSP 2014)*. Cologne, Germany, pp. 37–40.
- Birkholz, Peter, Susanne Drechsel, and Simon Stone (2019). “Perceptual Optimization of an Enhanced Geometric Vocal Fold Model for Articulatory Speech Synthesis.” In: *Interspeech 2019*. Graz, Austria, pp. 3765–3769.
- Birkholz, Peter and Dietmar Jackèl (2004). “Influence of Temporal Discretization Schemes on Formant Frequencies and Bandwidths in Time Domain Simulations of the Vocal Tract System”. In: *Interspeech 2004-ICSLP*. Jeju, Korea, pp. 1125–1128.
- Birkholz, Peter, Dietmar Jackèl, and Bernd J. Kröger (2006). “Construction and Control of a Three-Dimensional Vocal Tract Model”. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP’06)*. Toulouse, France, pp. 873–876.
- Birkholz, Peter and Bernd J. Kröger (2006). “Vocal Tract Model Adaptation Using Magnetic Resonance Imaging”. In: *7th International Seminar on Speech Production (ISSP’06)*. Ubatuba, Brazil, pp. 493–500.
- Birkholz, Peter, Bernd J. Kröger, and Christiane Neuschaefer-Rube (2011a). “Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds”. In: *First International Workshop on Performative Speech and Singing Synthesis (p3s 2011)*. Vancouver, BC, Canada.
- (2011b). “Model-based reproduction of articulatory trajectories for consonant-vowel sequences”. In: *IEEE Transactions on Audio, Speech and Language Processing* 19.5, pp. 1422–1433.
- (2011c). “Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis”. In: *Interspeech 2011*. Florence, Italy, pp. 2681–2684.
- Browman, Catherine P. and Louis Goldstein (1992). “Articulatory Phonology: An Overview”. In: *Phonetica* 49, pp. 155–180.
- Dang, Jianwu and Kiyoshi Honda (1994). “Morphological and Acoustical Analysis of the Nasal and the Paranasal Cavities”. In: *Journal of the Acoustical Society of America* 96.4, pp. 2088–2100.
- (1996). “Acoustic Characteristics of the Human Paranasal Sinuses Derived from Transmission Characteristic Measurement and Morphological Observation”. In: *Journal of the Acoustical Society of America* 100.5, pp. 3374–3383.
- Fant, Gunnar (1959). *Acoustic analysis and synthesis of speech with applications to Swedish*. Ericsson, Stockholm.
- Fant, Gunnar, Johan Liljencrants, and Qi guang Lin (1985). “A Four-Parameter Model of Glottal FLOW”. In: *STL-QPSR* 4, pp. 1–13.
- Ishizaka, K. and J. L. Flanagan (1972). “Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords”. In: *The Bell System Technical Journal* 51.6, pp. 1233–1268.
- Kane, J. and C. Gobl (2011). “Identifying regions of non-modal phonation using features of the wavelet transform”. In: *Interspeech 2011*. Florence, Italy, pp. 177–180.
- Klatt, Dennis H. and Laura C. Klatt (1990). “Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers”. In: *Journal of the Acoustical Society of America* 87.2, pp. 820–857.
- Narayanan, Shrikanth S., Abeer A. Alwan, and Katherine Haker (1997). “Towards Articulatory-Acoustic Models for Liquid Approximants Based on MRI and EPG data. Part I. The Laterals”. In: *Journal of the Acoustical Society of America* 101.2, pp. 1064–1089.
- Prom-on, Santhitam, Yi Xu, and Bundit Thipakorn (2009). “Modeling Tone and Intonation in Mandarin and English as a Process of Target Approximation”. In: *Journal of the Acoustical Society of America* 125.1, pp. 405–424.

Sondhi, Man Mohan and Juergen Schroeter (1987). “A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer”. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 35.7, pp. 955–967.