

VocalTractLab 2.0 User Manual

Peter Birkholz

April 26, 2012

Contents

1	Introduction	2
1.1	Purpose	2
1.2	Requirements, known issues	2
1.3	Download, Installation, Registration	2
2	Overview	3
2.1	Computational models	3
2.2	Graphical user interface (GUI)	3
3	Basic signal analysis (Signal page)	5
4	Vocal tract model analysis (Vocal tract page)	6
5	Time-domain simulation of vocal tract acoustics (Time-domain simulation page)	11
6	Gestural score (Gestural score page)	15
6.1	The concept of gestural scores	16
6.2	Editing a gestural score	16
6.3	Copy synthesis	17
7	Some typical uses	18
7.1	Analysis of the glottal flow for different supraglottal loads	18
7.2	Comparison of an utterance created by copy-synthesis with its master signal	18
7.3	Create and save a new vocal tract shape	19
7.4	Fitting the vocal tract shape to contours in an image	19
A	File formats	20
A.1	Speaker file (*.speaker)	20
A.2	Gestural score file (*.ges)	21
A.3	Segment sequence file (*.seg)	21
	References	22

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;
- analyze the time-varying distribution of pressure and volume velocity within the vocal tract during the synthesis of speech.

However, VTL is not a text-to-speech system. At the moment, connected utterances can only be synthesized based on gestural scores, as described in Sec. 6.

1.2 Requirements, known issues

VTL is currently developed for the Windows platform. It was tested with Windows XP, Vista, and 7. The minimum screen resolution is 1024x768, but higher resolutions are preferable. As the simulations are computationally intensive, a fast computer is recommended. For these reasons, tablet computers and netbooks are generally not suited to work with VTL.

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. Currently, there is the following issue: With Windows Vista and 7, you should use the “Aero design” for the desktop (this is the default). Otherwise, the model of the vocal tract, which is drawn using OpenGL, might not be displayed properly.

To run VTL on other platforms than Windows, e.g., Linux or Mac, we recommend to use a virtual machine, for example VirtualBox (www.virtualbox.org). 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 Download, Installation, Registration

The program 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 and start the program by calling “VocalTractLab2.exe”. The ZIP archive contains a couple of other data files, most of which are example files. The only essential file next to the executable is “JD2.speaker”, which is an XML file that defines the default speaker (see Sec. A.1) and is loaded automatically when VTL is started. After downloading the program, it has a limited functionality. To get access to all functions, please register your copy of VTL. Just write a short email containing your name, city, and institution to peter.birkholz@vocaltractlab.de with the request to register. In response, you will receive a personalized registration file activating all the functions described in this manual. Note that the response may take a few days, as it is done manually. If you have a registration file for a previous version of the software, you can also use it for the new version and don’t need to register again. Just copy the file “registration.txt” into the folder with the new version.

2 Overview

2.1 Computational models

VTL implements various models, for example 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. It has currently 22 degrees of freedom (vocal tract parameters) that control the shape and position of the model articulators. The model was originally developed by Birkholz (2005) and later refined by Birkholz, Jackèl, and Kröger (2006), and Birkholz and Kröger (2006). The currently implemented version of the vocal tract model was even further improved and will be described in a forthcoming publication. For the simulation of acoustics, the area function of the vocal tract is calculated, i.e., the variation of the cross-sectional area along the center line of the model. 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, there are two approaches 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 geometrical model by Titze (1989), 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). These models can be individually parameterized and exchanged for the acoustic simulation. 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 Stevens (1998) is used.

Coarticulatory effects of the vocal tract are modeled using a dominance model. In this model, for the formation of a certain consonantal constriction, each articulator (i.e., vocal tract parameter) has an individual dominance value that defines how “important” the articulator is for that particular constriction. For example, for an apico-alveolar closure as in /t/, the parameters controlling the position of the tongue tip have a high dominance (they are important), and the parameter controlling the height of the larynx has a low dominance (it is less important). Hence, the height of the larynx is more influenced by the vocal tract shape of the context vowel(s) than the tongue tip position. A more detailed description of this mechanism is given by Birkholz and Kröger (2006).

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. 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).

2.2 Graphical user interface (GUI)

Fig. 1 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 title bar shows for whom the copy of the program is registered. The menu bar has the four menus “File”, “Export”, “Synthesis models”, and

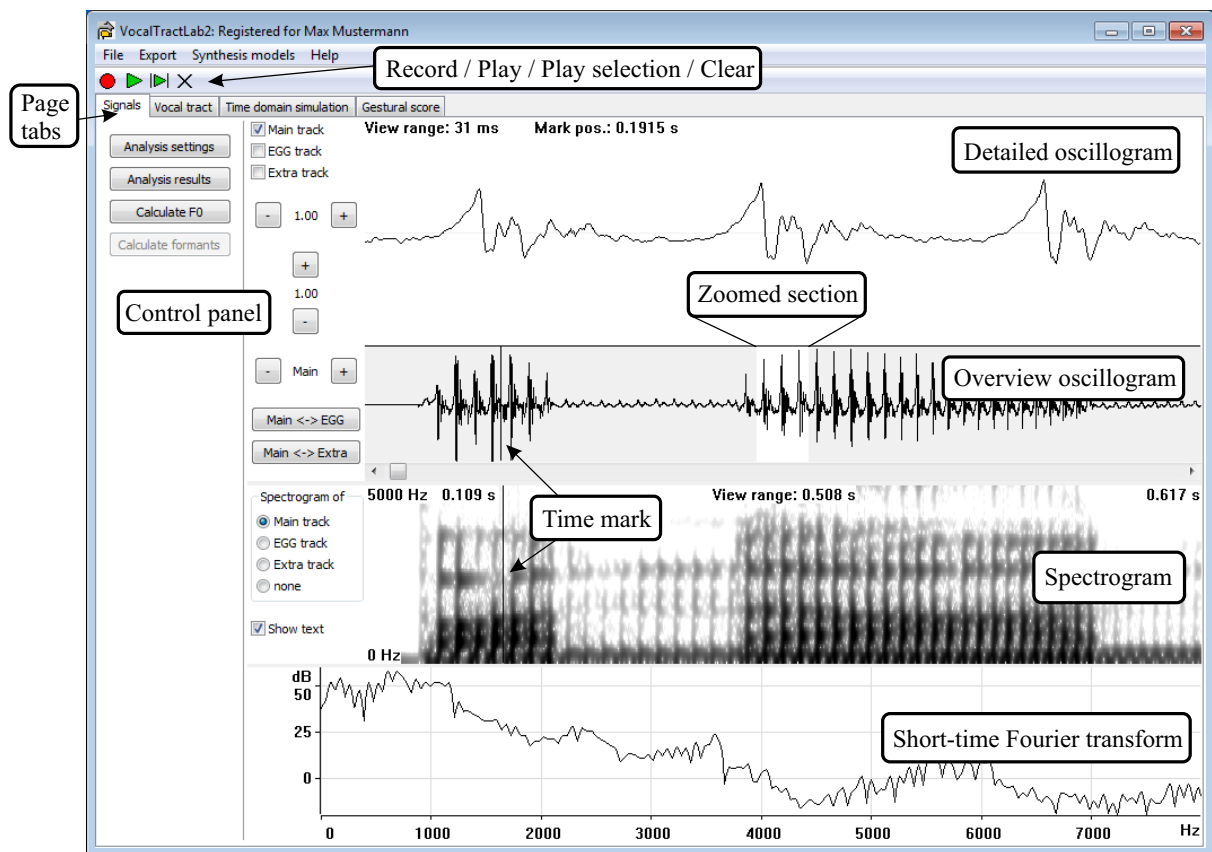


Figure 1: Signal page.

“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. 1, is meant for the acoustic analysis of speech signals.
- The *vocal tract page* is meant for the analysis of the vocal tract model and the synthesis of vowels in the frequency domain.
- The *time-domain simulation page* is meant for the analysis of vocal tract acoustics in the time domain.
- The *gestural score page* is meant for the manual creation of gestural scores and the copy-synthesis of natural utterances.





The following four sections will explain the basic functionality of the pages. In general, each of the pages consists of a main region that contains different displays, and a control panel at the left 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 “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 sampled 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 meant 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 different 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 22050 Hz and a quantization of 16 bit. The menu “File” has five items to load and save sampled 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 signal 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 22050 Hz, the sampling rate is converted automatically.

There are four displays in the main part of the signal page (cf. Fig. 1): the detailed oscillogram, the overview oscillogram, the spectrogram, and the short-time Fourier transform. With 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  +  or  + . 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. 2) 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  + . 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  + . The green arrow between the two vertical lines or the key combination  +  plays the main track signal in the selected time range. This signal part is played in a loop.

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

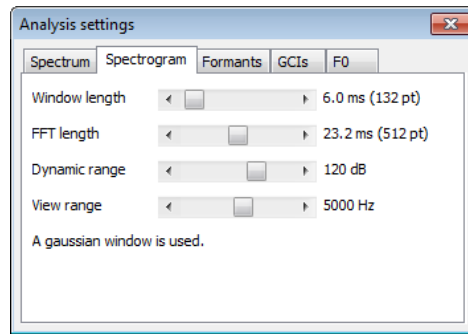


Figure 2: Dialog for analysis settings.

- “Analysis settings” opens the dialog shown in Fig. 2. This dialog allows setting the parameters used for the calculation of the spectrum in the bottom display, for the spectrogram, and the fundamental frequency (F0) contour.
- “Calculate F0” calculates the F0 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).
- “Analysis results” opens a non-modal dialog where the F0 at the time of the mark in the oscillogram/spectrogram is displayed in semitones (relative to the musical note C_0 with 16.352 Hz) and Hz.

4 Vocal tract model analysis (Vocal tract page)

Fig. 3 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. 4). This dialog is opened automatically when the program is started. If it has been closed, it can be shown again by selecting the menu item “Synthesis models → Vocal tract model” or the button “Show vocal tract” in the control panel. The parameters of the model can be changed by dragging the yellow control points with the LMB. 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. Seven vocal tract parameters, which cannot be controlled with the control points, are adjusted using the scrollbars below the vocal tract display. Display options for the vocal tract model can be found at the bottom of the dialog. Furthermore, you can load an image from a GIF 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.

The area function corresponding to the current vocal tract shape is shown in the top left display of the vocal tract page. In this image, the area functions of the sinus piriformis and the nasal cavity are shown, too (the sinus piriformis are currently not considered in the acoustic simulation of the vocal system by default). 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.

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 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). The formant frequencies are automatically determined and marked by vertical dashed lines.

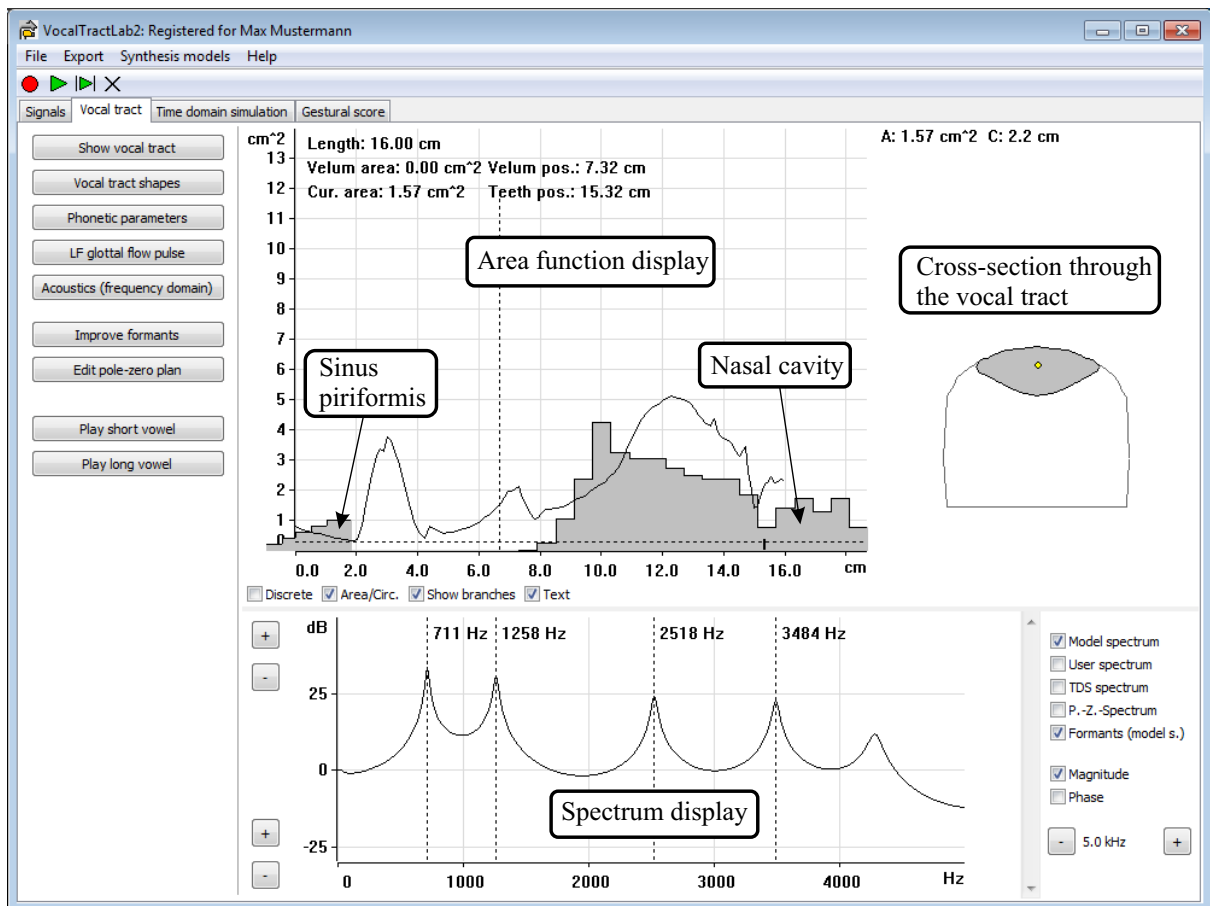


Figure 3: Vocal tract page.

Note that the formant determination is not reliable when there are zeros (anti-formants) in the transfer function, for example when the velo-pharyngeal port is open. A left-click in the spectrum display opens the dialog in Fig. 5. Here you can select the kind of model spectrum to display:

- “Transfer function U/U ” is the volume velocity transfer function between the glottis and the lips.
- “Transfer function P/U ” is the complex ratio of the radiated sound pressure and the volume velocity at the glottis.
- “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” is the spectrum of the sound pressure that would be measured in front of the mouth, when the vocal tract model would be 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

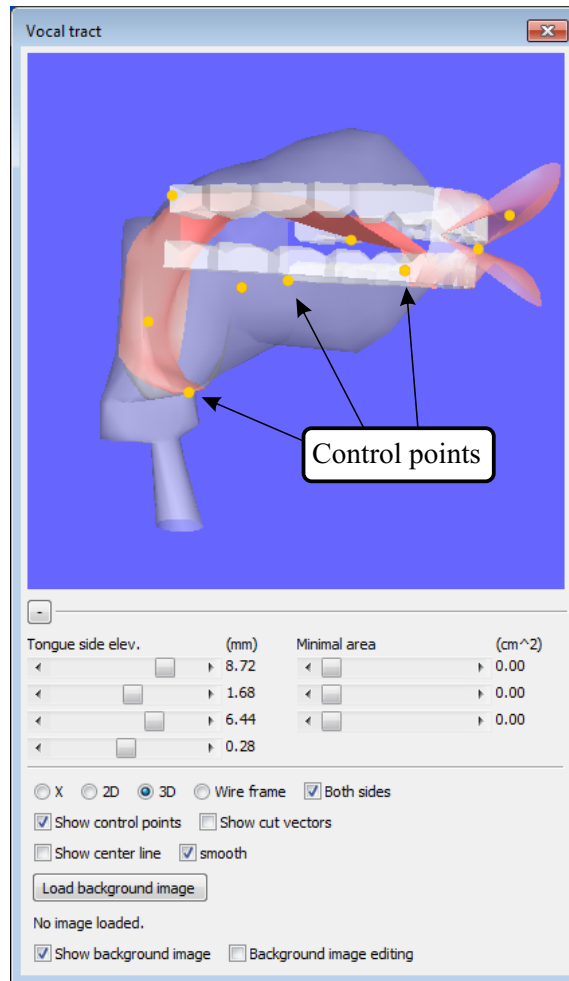


Figure 4: Vocal tract dialog.

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).

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 regard the loss mechanisms that are considered. They can be changed in the dialog shown in Fig. 6, 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 leave it at the default value.

With the buttons “Play short vowel” and “Play long vowel” 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. 7, 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 vocal tract model comes with a number of predefined shapes for German vowels and typical consonantal constrictions at different places of articulation. These shapes are used as articulatory targets when

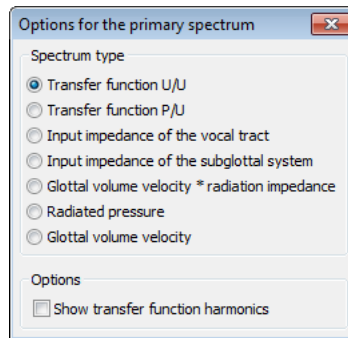


Figure 5: Options for the model spectrum to display.

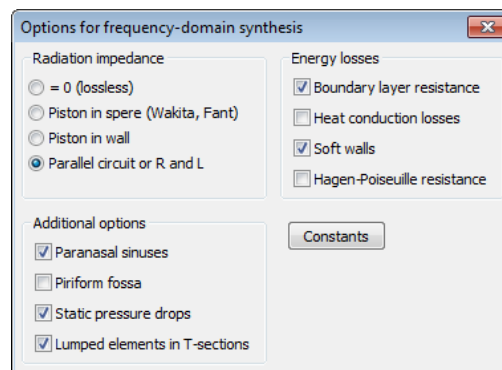




Figure 6: Options for the acoustic simulation in the frequency domain.

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. 8, 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 and dominance values are shown at the right side. The concept of the dominance values was briefly introduced in Sec. 2.1 and is discussed in detail in Birkholz and Kröger (2006). How the vocal tract shapes were obtained is also described in Birkholz and Kröger (2006).

The names chosen for the vocal tract shapes of vowels are the corresponding SAMPA symbols. For each of the three diphthongs /aU/, /aI/ and /OY/, there is one initial and one final shape of the vocal tract in the list. 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 “-clo” for a full closure, “-cons” for a critical constriction, and “-lat” for a lateral constriction. This naming scheme is later used in gestural scores to distinguish consonantal gestures with respect to the primary articulator. 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.

The button “Improve formants” in the control panel allows the automatic adjustment of the vocal tract shape such that the first three formants in the transfer function approach specific values given by the user. An 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. This may be interesting, for example, when

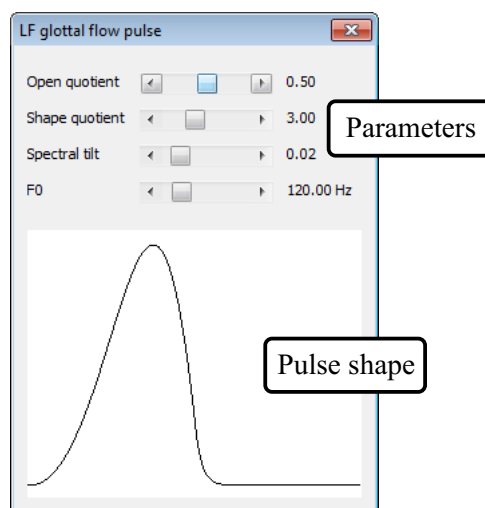


Figure 7: LF glottal flow pulse dialog.

you want to adapt the formants of a vowel to the realization of that vowel in a different language or dialect. Please note that this is not a method for full Acoustic-to-Articulatory Inversion, because the vocal tract parameter space is only searched in the vicinity of the current vocal tract configuration.

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. 9. There are seven parameters, which can be changed by scroll-bars. 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.

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. 10, click the button “Edit pole-zero plot” 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”. 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.

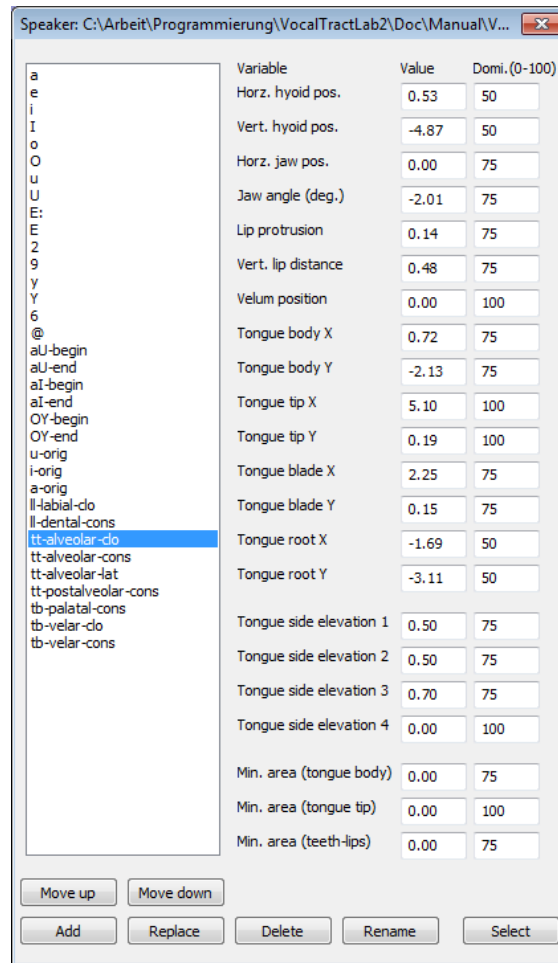


Figure 8: Vocal tract shapes dialog.

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

Fig. 11 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) distribution in the vocal tract changes 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 14 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 and displayed by four circles. The colors of the tube sections indicate the magnitude of the selected acoustic 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 or the volume velocity (flow). In Fig. 11, the flow is visualized. Above the area function display is the space-time graph. Here, the selected variable is shown as a curve. The black curve shows the pressure or flow 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. 11, the upper glottis section is selected. In the upper display

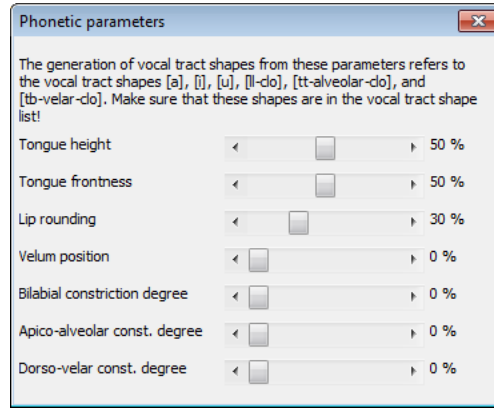


Figure 9: Phonetic parameters dialog.

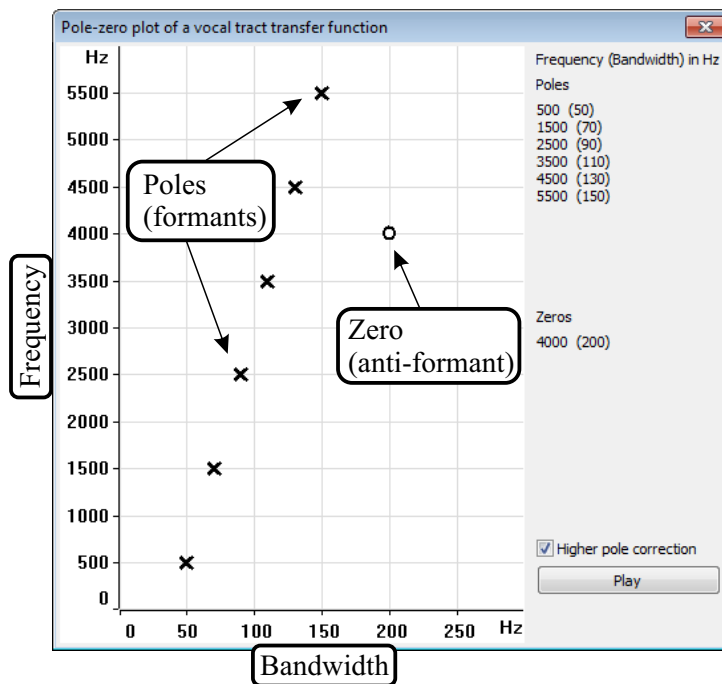


Figure 10: Pole-zero plot.

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. 11 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

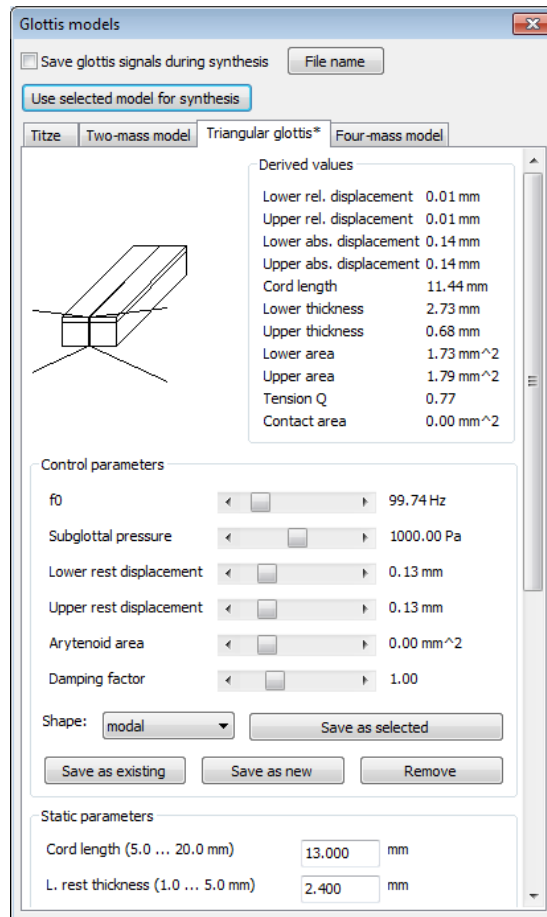


Figure 12: Glottis dialog.

VTL has implemented different models of the vocal folds that can be used for acoustic synthesis in the time domain. These models are managed in the glottis dialog shown in Fig. 12, which can be called with the button “Glottis dialog” in the control panel or from the menu “Synthesis models → Vocal fold models”. Currently, four vocal fold models are implemented: the geometric model by Titze (1989), the classical two-mass model by Ishizaka and Flanagan (1972), the modified two-mass model by Birkholz, Kröger, and Neuschaefer-Rube (2011c), and a four-mass model. The four-mass model is not intended to be used yet. 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 pressed. 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 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 parameters or shapes, you must save the speaker file by pressing **[F2]** or selecting “File → Save speaker” from the menu.

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.

6 Gestural score (Gestural score page)

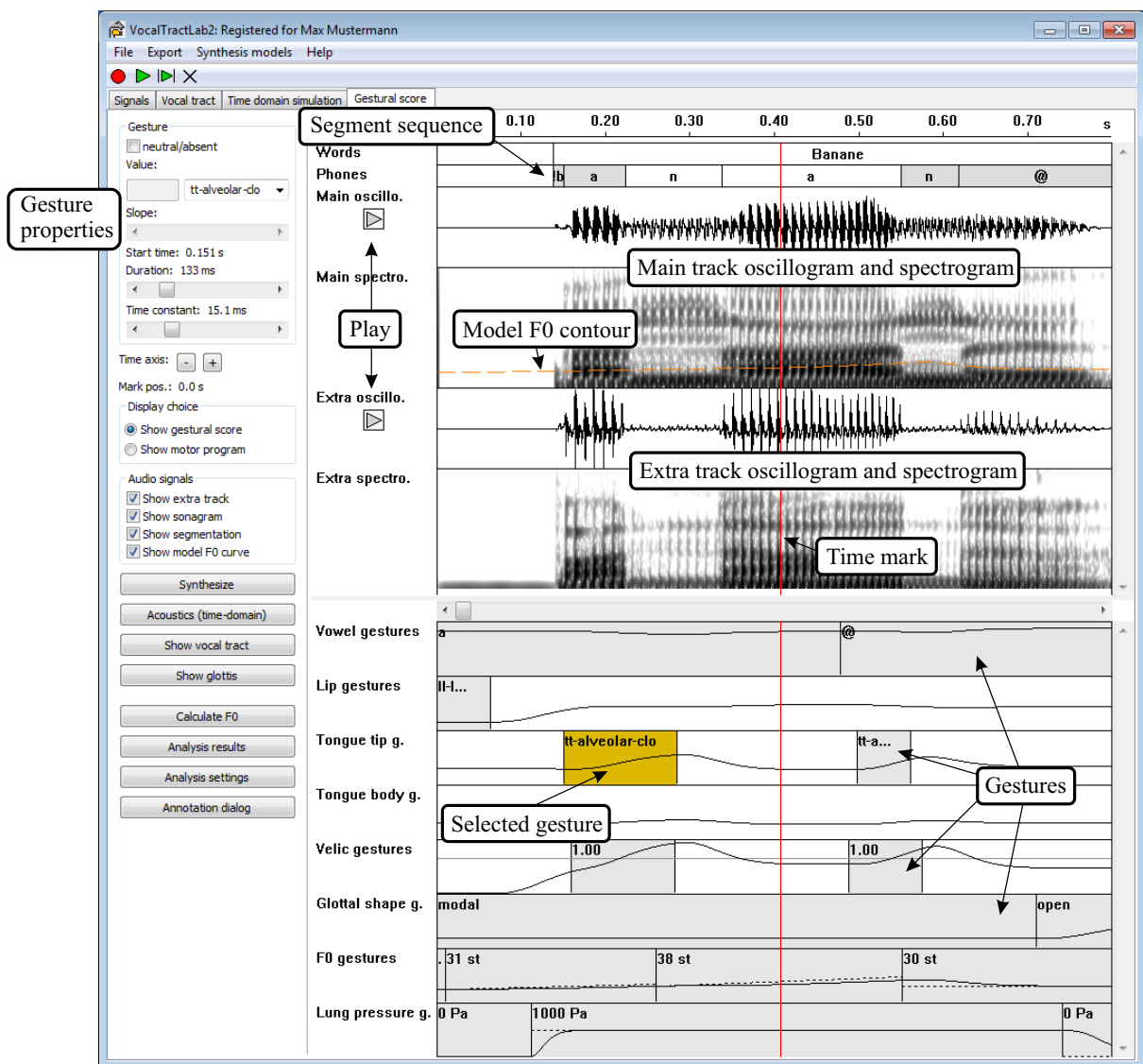


Figure 13: Gestural score page.

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. 13). 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. 13. 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 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). In Fig. 13, for example, the score starts with the glottal shape “modal” for modal phonation, and approaches an open state with the gesture “open” towards the end.

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 word [banan@] in Fig. 13, there is one long vowel gesture for [a], superimposed with a lip gesture for the bilabial closure for [b] and a tongue tip gesture for the apico-alveolar closure of [n]. A second tongue tip gesture for the second [n] starts at about the same time as the basic diphthongal movement from [a] to [@]. Parallel to the tongue tip gestures are two velic gestures to open the velo-pharyngeal port during the oral closures. 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. The degree of participation of the individual articulators in the realization of a certain consonantal constriction is defined by the dominance values for that shape (cf. Sec. 4).



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 active gesture for the bilabial closure is followed by a neutral gesture that simply represents the movement back to the underlying vocalic target for [a]. 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. 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. To find the “right” time constant for a gesture is somewhat tricky, but a value of 15 ms for supraglottal gestures has proven to give satisfactory results in most cases. The value of a gesture is either a numeric value (the subglottal pressure in Pa for pressure gestures, the F0 in semitones 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.

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:

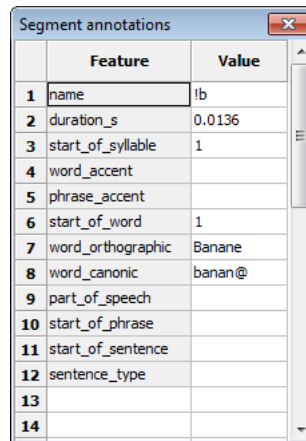
-  +  simultaneously increases the length of the gestures at the time mark in all tiers by a small amount. In this way you can “stretch” the gestural score at the time of the mark, for example to increase the length of the phone at that time.
-  +  simultaneously decreases the length of the gestures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

6.3 Copy synthesis

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 utterances 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. 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 calling the context menu in the segment row and move the borders between segments by dragging them with the LMB. A double-click on a segment opens the segment annotation dialog shown in Fig. 14, where the properties of a segment can be changed. For a description of the properties please refer to Sec. A.3.

Which parts of the display are shown in the top right part of the page can be controlled in the button group “Audio signals” in the control panel. The checkbox “Show model F0 curve” allows showing the





	Feature	Value
1	name	lb
2	duration_s	0.0136
3	start_of_syllable	1
4	word_accent	
5	phrase_accent	
6	start_of_word	1
7	word_orthographic	Banane
8	word_canonic	banan@
9	part_of_speech	
10	start_of_phrase	
11	start_of_sentence	
12	sentence_type	
13		
14		

Figure 14: Segment annotation dialog.

model F0 curve as a red dashed line in the spectrogram of the main track to compare it with the measured F0 contour of the master signal in the extra track.

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. 11. The time signal of the selected section and the selected acoustic 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.
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 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.

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 “banane-synth.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, uncheck the checkbox “Show animation”, press “OK”, 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 “banane-orig.wav” to the *extra track* with the menu item “File → Load WAV”.
5. Load the segment sequence “banane-orig.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. Your screen should look similar to Fig. 13 now.
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.

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” 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 [u] 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” in the control panel of the vocal tract page to hear the corresponding vowel. When the contours coincide well, you should hear an [u].
7. Click the button “Vocal tract shapes” to open the vocal tract shapes dialog and double-click on the shape “u-orig”. the parameters of this shape were manually adjusted to coincide as well as possible with the background image.

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.

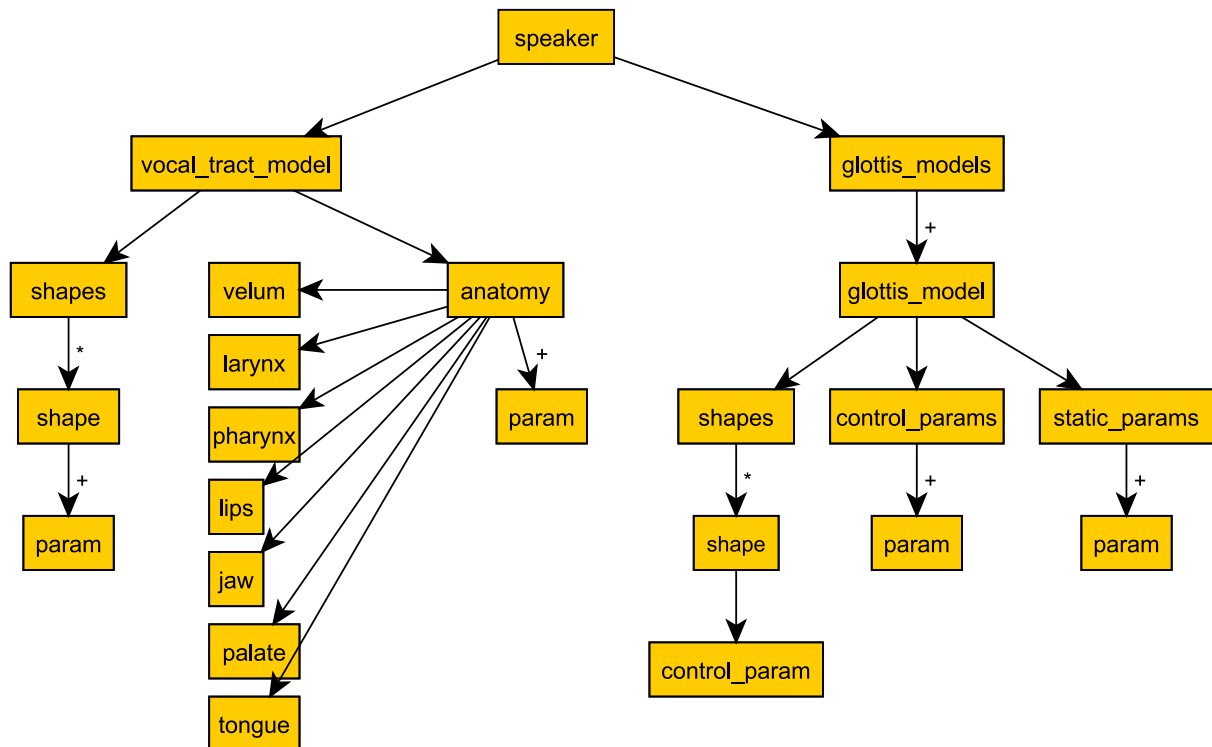


Figure 15: DTD tree of the speaker file.

The document type definition (DTD) tree in Fig. 15 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. 16 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.

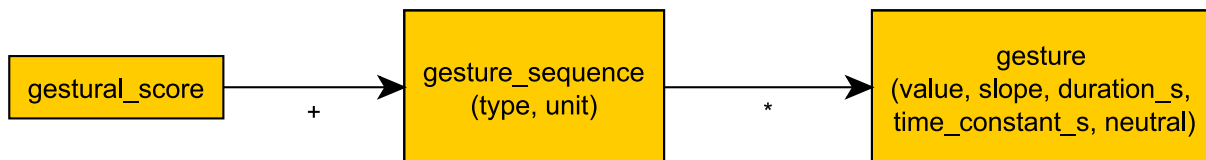


Figure 16: DTD tree of the gestural score file.

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.

Acknowledgments

I thank Ingmar Steiner, Phil Hoole, Eva Lasarczyk, Véronique Delvaux, Yi Xu, Santitham Prom-On, and Lucia Martin for testing the software, proofreading this manuscript and general feedback. Parts of the research leading to the development of VocalTractLab were funded by the German Research Foundation (DFG), grant JA 1476/1-1.

References

- 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.
- 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*. Jeju Island, 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.
- 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, Kenzo and James 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.
- 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* ASSP-35.7, pp. 955–967.
- Stevens, Kenneth N. (1998). *Acoustic Phonetics*. The MIT Press, Cambridge, Massachusetts.
- Titze, Ingo R. (1989). “A Four-Parameter Model of the Glottis and Vocal Fold Contact Area”. In: *Speech Communication* 8, pp. 191–201.