CHAPTER 10

ARTICULATORY SPEECH SYNTHESIS TOOLBOX

10.1 INTRODUCTION

The aim of this chapter is to provide a flexible, articulatory speech synthesis toolbox, called ARTM, for articulatory speech synthesis model. One feature of this toolbox is to synthesize speech by matching the vocal tract parameters for a specified set of formant characteristics, called the target formants. A simulated annealing procedure is used to achieve the optimization. The derivation of the acoustic equations that include the subglottal system, the glottal impedance, the turbulence noise source, and the nasal tract with sinus cavities for the articulatory synthesizer is provided in Appendix 11. This toolbox can be useful for speech modeling, analysis, and synthesis.

The toolbox is designed with interfaces that provide for numerical specification of parameters and sliders that allow parameter adjustments. A transmission-line circuit model of the vocal system, which includes the vocal tract, the nasal tract with sinus cavities, the glottal impedance, the subglottal tract, the excitation source, and a turbulence noise source, are provided. A digital time–domain approach is used to simulate the dynamic properties of the vocal system as well as to improve the quality of the synthesized speech.

There are two components of the toolbox: the formant_track folder and the artm folder. These folders can be installed in the same or separate subdirectories of MATLAB. The use of these toolboxes is described in the sections that follow.

The analysis determines the articulatory parameters from the acoustic speech waveform. The algorithm that is used is known as simulated annealing, which is constrained to avoid non-unique solutions and local minimal problems. The articulatory-to-acoustic transformation function and the boundary conditions for the articulatory parameters determine the constraints. The cost function is defined as a percentage of the weighted least-absolute-value error between the first four formant frequencies of the articulatory model (the model formants) and the first four formant frequencies determined from speech analysis (the target formants). A 1% error criterion is both practical and achievable.

In summary, this articulatory speech synthesis toolbox works as follows. The user analyzes a speech file, such as a word or a sentence to determine the target formants for a set of frames defined for the speech file. This preliminary task is performed with a special toolbox that is derived from the formant/pitch tracking option in the analysis toolbox described in Chapter 2. In this chapter, the target formants are derived using the formant_track folder, which is described later. Once the target formant tracks are obtained, the user then opens the articulatory speech synthesis toolbox and loads either a formant track file or an articulatory parameter file. The loading of an articulatory parameter vector is explained more fully below. After a formant track file is loaded, a set of articulatory parameters (a vector of parameters) is determined by minimizing the error between the target formants and the model formants for each frame. The articulatory vector determines the vocal track shape for the articulatory speech synthesizer. Typically, the user saves the

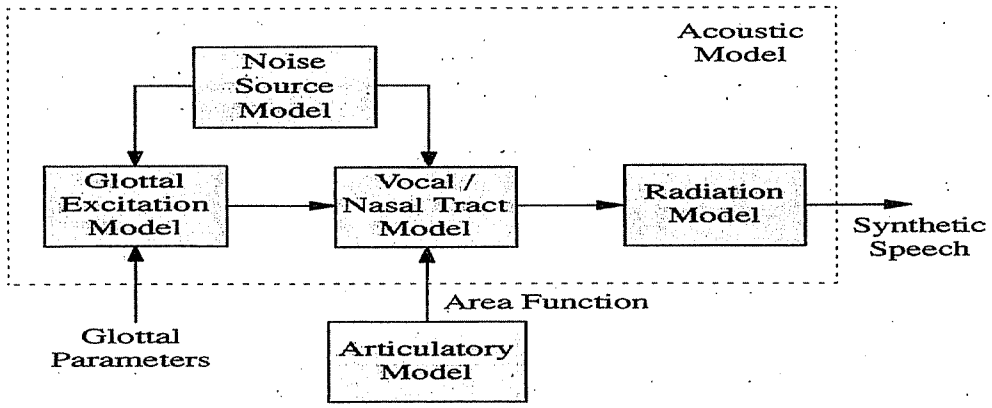


FIGURE 10.1 Overview of the articulatory speech synthesizer.

articulatory vector to a file, which can be loaded at a later time if desired. An example is given later. After the articulatory parameter vector has been determined, the user then specifies the excitation waveform using the LF parameters. The excitation can be voiced, unvoiced, or a mixture of voiced and unvoiced excitation. The voiced excitation can include jitter and shimmer, as well as aspiration and turbulence. After the excitation is specified, the speech is synthesized using the excitation to drive the vocal tract determined by the articulatory parameter vector. The speech synthesizer includes options to display an animated movie of the vocal tract used for speech synthesis, to display the vocal tract parameters, to play the synthesized speech, as well as other features. The toolbox provides the user with a method for observing the relationship between speech features (formants) and vocal tract shape.

While Appendix 11 provides the theoretical background for the articulatory speech synthesizer, we will from time-to-time give brief descriptions of important features of the system. Figure 10.1 shows an overview of the articulatory synthesizer, which consists of an articulatory model, an acoustic model, and an excitation source. The acoustic model is outline in Figure 10.2 and is described in more detail in Appendix 11. The major advantages of the articulatory approach to speech synthesis are that (1) the model is directly related to human speech production, therefore the model parameters vary slowly, and are easily interpolated; and (2) source-tract interaction is modeled in a natural manner. However, difficulties exist, for example (1) it is difficult to derive an accurate articulatory model, including an accurate representation of the losses; (2) the estimation of the model parameters by inverse filtering is difficult due to local minima and the solution can be non-unique; and (3) it is still difficult to model unvoiced sounds, although we have improved this feature (see Appendix 11). Nevertheless, our articulatory speech synthesizer is complete and user friendly.

There are four phases or steps in the use of the articulatory synthesizer: analysis, speech inverse filtering, excitation specification, and synthesis. The analysis phase extracts the target formants for the speech file by determining the formant tracts of the target speech signal. The user then marks the target formant trajectories at desired intervals (frames). This marked formant trajectory file is saved as the target formant file. Next, speech inverse filtering is performed to determine the articulatory model parameters. This is accomplished using a simulated annealing algorithm to minimize the distance (error) between the target formants and the model formants (see Appendix 11). The user then specifies the type of excitation that is to be used for synthesis. Finally, the speech synthesis is performed. The synthesis option provides an animated display of the vocal tract configuration used for synthesis as well as displays of other model parameters. A summary of the articulatory synthesizer is given in Figure 10.3. The vocal tract area function at the top right is for the vowel /AA/.

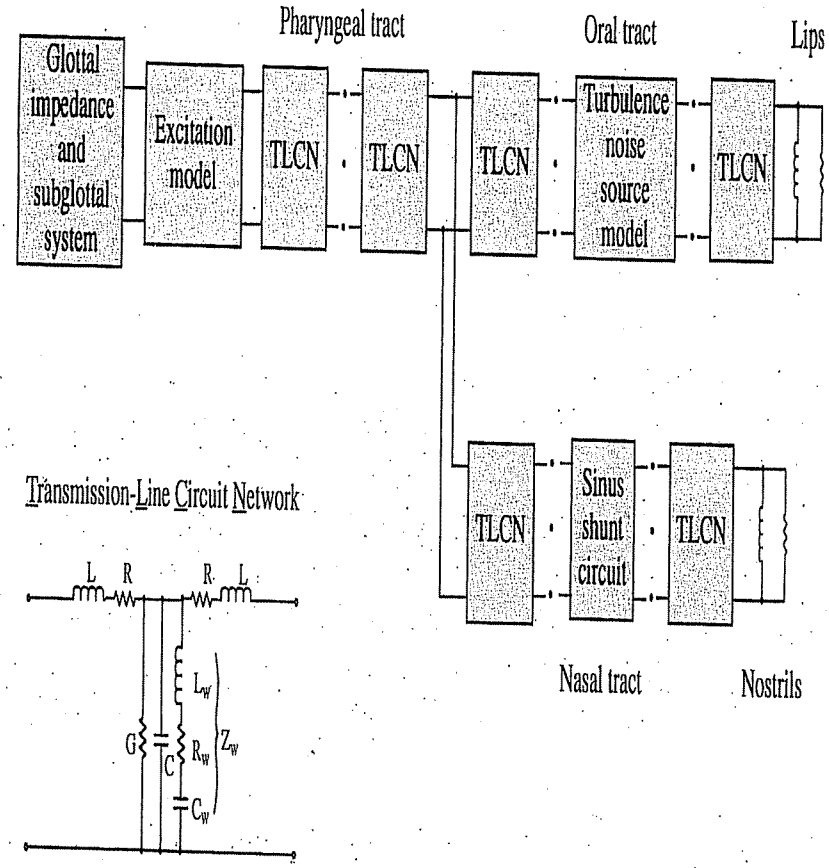


FIGURE 10.2 The acoustic model.

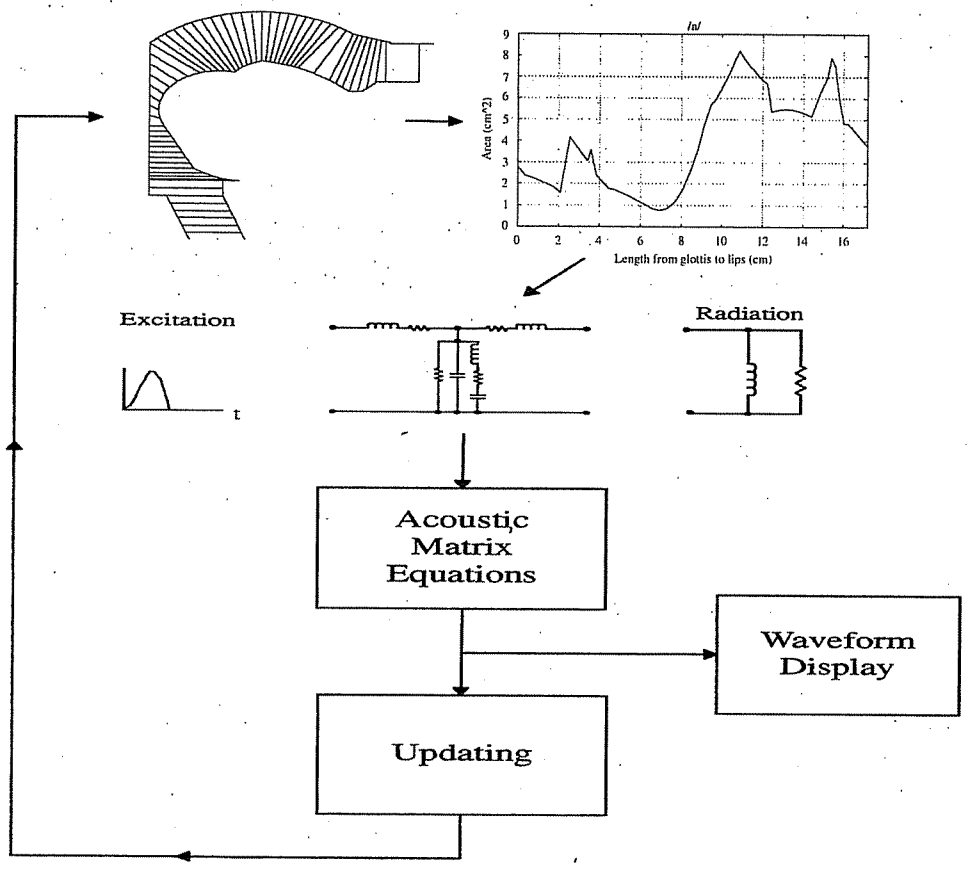


FIGURE 10.3 The steps in articulatory speech synthesis.

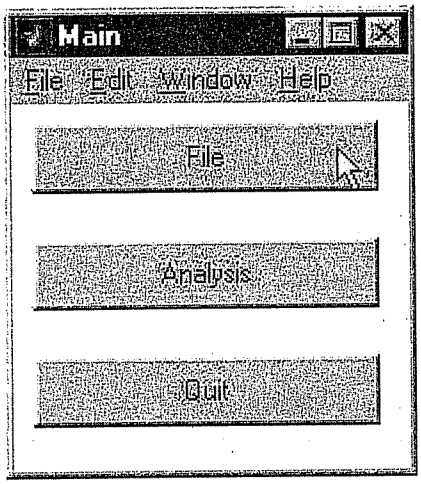


FIGURE 10.4 Main window for estimating the target formant tracks for a speech file.

10.2 DETERMINING THE TARGET FORMANTS

The target formants for a speech file are determined using the formant_track toolbox. As for previous toolboxes, start MATLAB, change directory to formant_track, and type main. The window shown in Figure 10.4 appears, which is similar to the main window for the analysis toolbox in Chapter 2.

Next, press the File button to open the file window shown in Figure 10.5. Pressing the Load button allows the user to load an ASCII speech file, such as b.dat or be_seg.dat, both of which are in the data folder in the formant_track toolbox. Figure 10.6 shows the b.dat file as the input signal file. Once the desired speech file is loaded, the user can play the file if desired or the option can be canceled.

After the desired speech file is loaded, the user returns to the main window and presses the Analysis button, which opens the analysis window shown in Figure 10.7. The options available are to calculate four formant contours (tracks), after which the user can save the formant contours by pressing the Save button. The file name is specified as *form.for, where the * is to be replaced by a name specified by the user, such as be_. The formant contour file name then becomes be_form.for. Such a file is already available in the data folder within

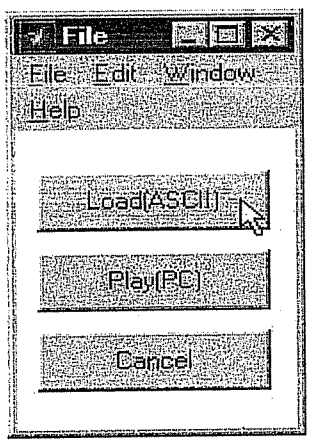


FIGURE 10.5 File window for loading a speech file.

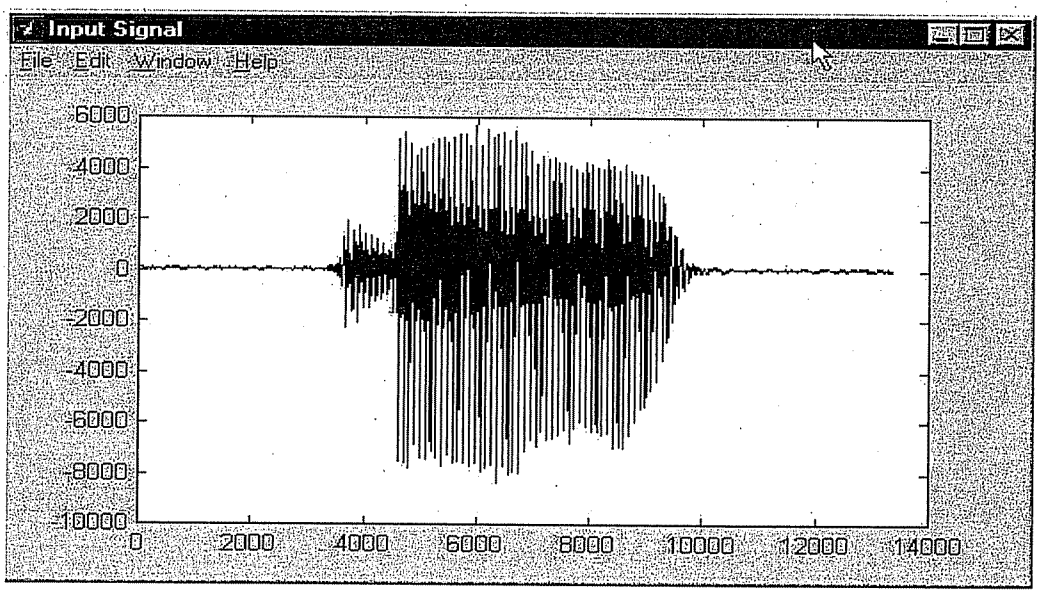


FIGURE 10.6 Input signal file.

the formant_track toolbox. Pressing the Cancel button performs the customary function. Pressing the Formant Contours button calculates the first four formant contours for the loaded speech file, which in this case is b.dat, shown in Figure 10.6. While the formant contours are being calculated several short message windows are displayed, as is done in the analysis toolbox in Chapter 2. Upon completion of the calculations, the formant contours are displayed as shown in Figure 10.8.

The saved formant tracks are used in the articulatory speech synthesizer toolbox, which is described next.

10.3 ARTICULATORY SPEECH SYNTHESIS TOOLBOX

After installing the articulatory speech synthesis toolbox in a subdirectory, start MATLAB, change directory to the subdirectory, and type main. Two windows appear. One is a menu window that contains the buttons shown in Figure 10.9. The other, shown in Figure 10.10,

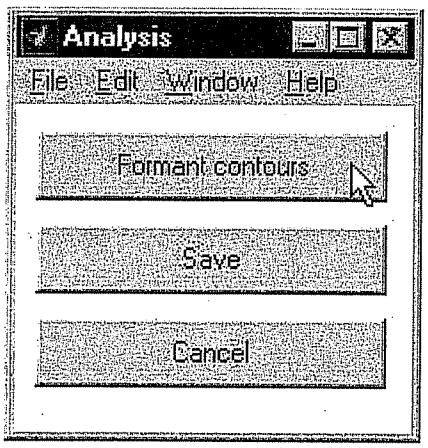


FIGURE 10.7 The analysis window.

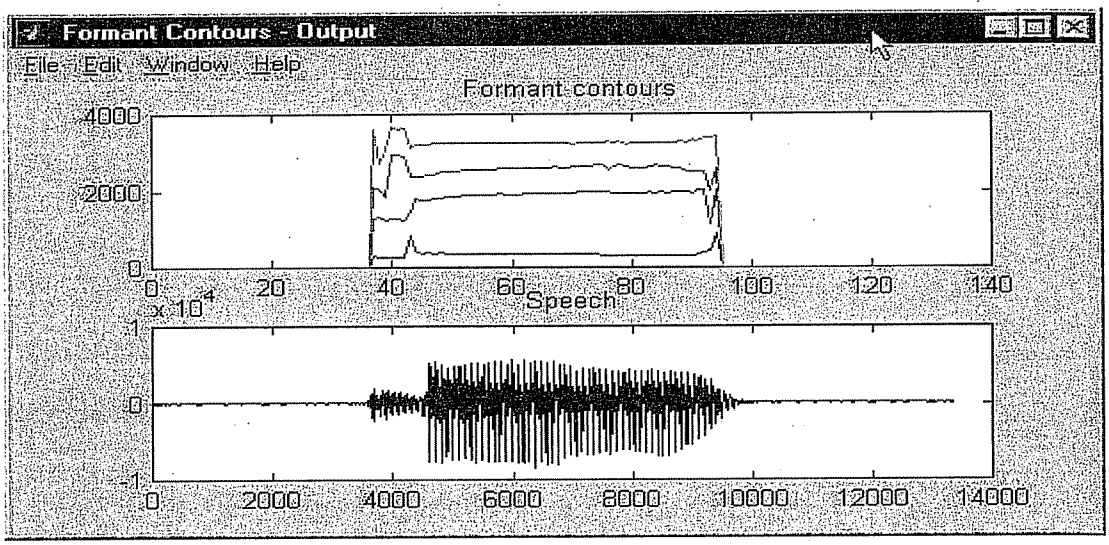


FIGURE 10.8 Formant contours and speech file.

is a canvas window that is used to display graphics for the various options. The user is not to close these two windows until he or she is ready to exit the toolbox.

There are six buttons (options) available in the main menu window: File is to load or to save a file, Shape is to set the articulatory model parameters, Optimize is to start the optimization process, Excitation is to set the desired excitation parameters for the excitation waveform, Synthesize is to start the synthesis process, and Quit is to quit and exit the toolbox. Generally these buttons are pressed in a left to right sequence.

10.3.1 The Load and Mark Option

The load feature allows the user to load one of three types of files: a formant tracks file (determined using the formant_track toolbox), an articulator parameter file, or a Fant area file (as shown in Appendix 5). The latter option is disabled at this time. See the comments that appear when this option is selected for additional information. The load articulatory parameter file is explained later. We start with the load formant tracks option. This option opens the data folder within the articulatory speech synthesis toolbox showing the files named *form.for, which includes be_seg_form.for. This latter file is the formant track file for the speech file be_seg.dat, which is a short segment of the b.dat file that starts just following the termination of the plosive /B/. The original speech file can be viewed using the formant_track toolbox by loading the be_seg.dat file. Load the be_seg_form.for file. The windows shown in Figures 10.11 and 10.12 appear. Figure 10.11 displays the four formant tracks for the be_seg_form.for file, while Figure 10.12 shows a menu window with five options. The primary options are contained in the pull-down menu, which are Mark and Add Mark. The term mark is used to designate the operation of marking temporal locations

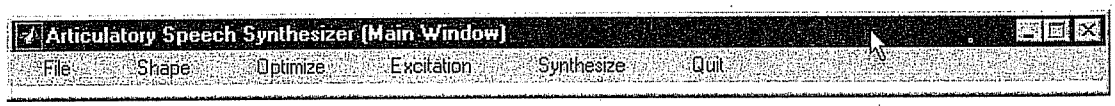


FIGURE 10.9 The articulatory speech synthesizer menu window.

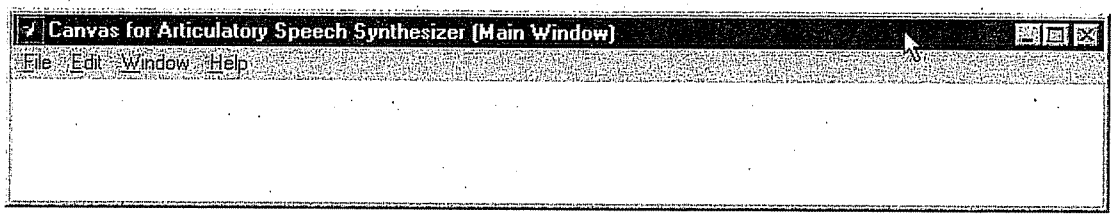


FIGURE 10.10 The canvas window.

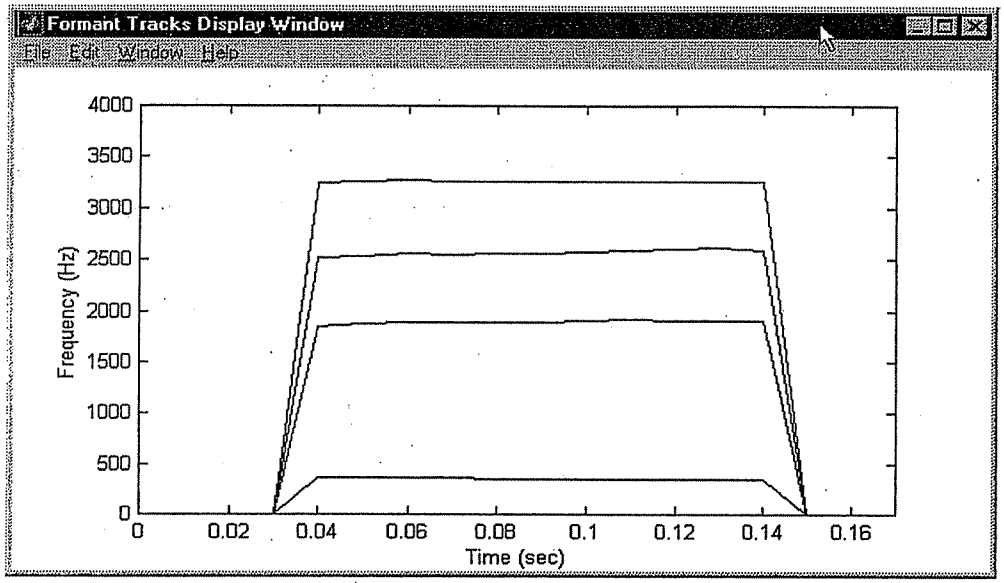


FIGURE 10.11 Formant tracks display window.

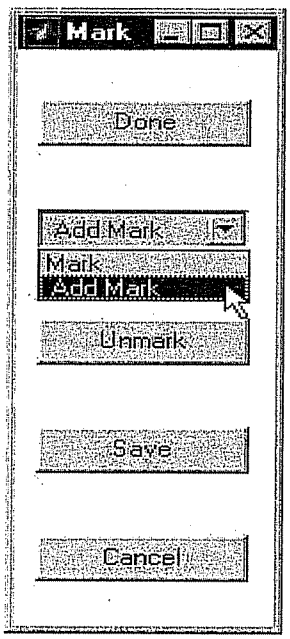


FIGURE 10.12 The formant track mark menu.

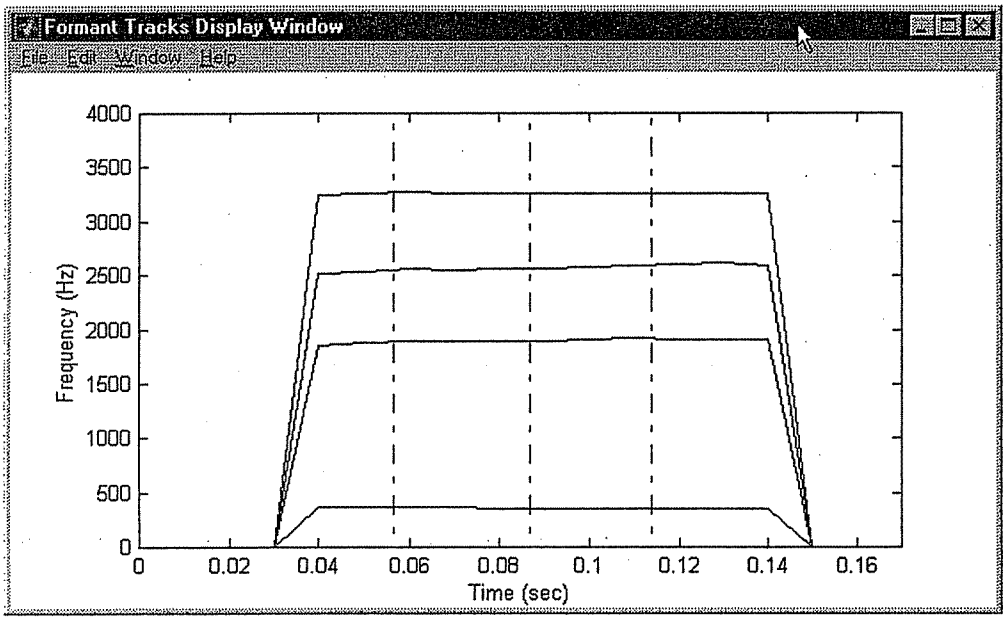


FIGURE 10.13 An example of a formant track file that has been marked using the add mark option.

for frame boundaries on the formant tracks. An example of this is shown in Figure 10.13, where the add mark option was used to add three frame boundaries on the formant tracks. This was accomplished by selecting the add mark menu option, then moving the cursor to the formant tracks display window, whereupon the cursor becomes a cross hair. The user places the cross hair at the desired location and clicks the left mouse button, a vertical frame boundary line is drawn at that location. This sequence of steps was repeated two more times to draw the three frame boundaries as shown in Figure 10.13. The purpose of the segmentation of the target formant tracks into boundaries is to facilitate the calculation of the shape of the vocal tract. This calculation minimizes the distance between the target formants and the formants of the articulatory model. The user can Unmark the file by pressing the unmark button. The operation of unmarking is complete; that is, all boundary marks are removed. The user must then initiate the add mark process again. Once the number of desired frame boundaries are marked at the desired locations, then save the file for future use by pressing the Save button. The save option opens the data folder and allows the user to name the marked file using the format *fmrk.mrk, where the * is to be replaced with a user selected name, such as be_seg_. After saving the marked frame boundary file, then press the Done button. The Done button is pressed only if a formant track file has been marked. The Done button can be pressed without saving the marked file. The marked file is retained in memory for the next phase of the toolbox. If a formant track file is not marked and the user wishes to exit this phase of the software, he or she can press the Cancel button. Finally, the purpose of the mark file option is to load a previously marked formant track file, such as be_seg_fmrk.mrk, which is the file shown in Figure 10.13 that was constructed using the add mark option and saved as described previously. Once such a previously marked file has been loaded, press the Done button to continue. The user can also unmark a loaded marked file. The original marked file is not destroyed, since it is retained in disk storage. However, all of the marked frame boundaries are removed from the software memory workspace and the user must add the desired frame boundaries in the manner described. In other words, it is as though the user had loaded the original formant track file, such as the be_seg_form.for file in the first place. So there is no advantage to unmarking a previously saved marked file.

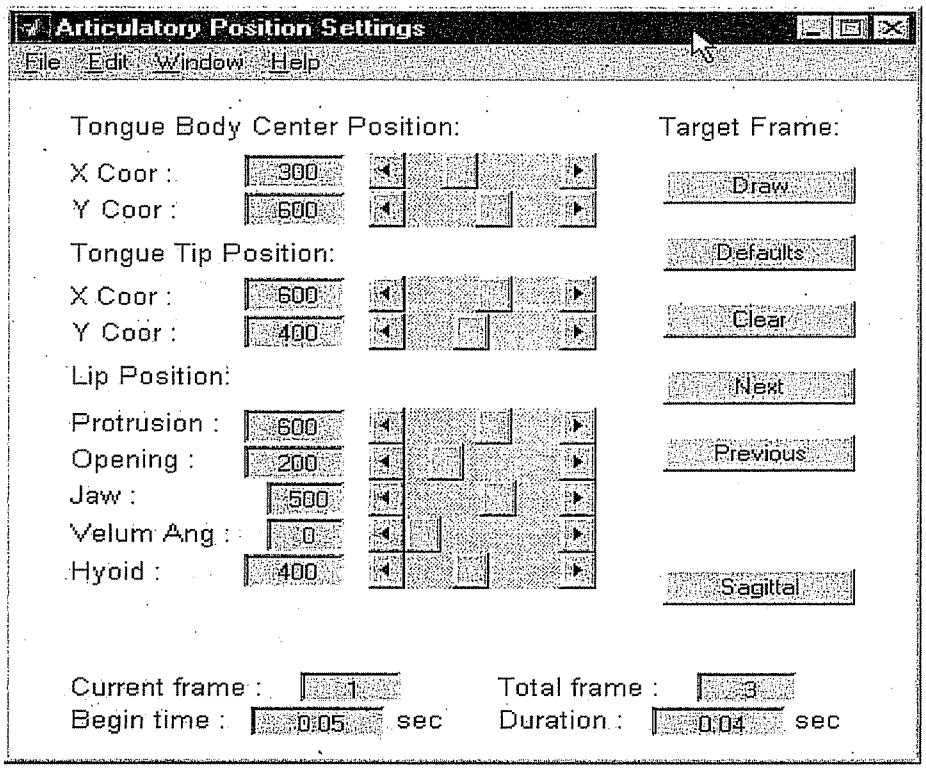


FIGURE 10.14 Articulatory positions settings window.

The number of frames to be marked on a formant track file is not critical. For example, the sentence, "We were away a year ago," is shown in Appendix A11-D for two different speakers marked at 26 and 22 locations for the entire sentence. The synthesis was quite satisfactory with so few frames. The word be can be synthesized with 3 frames.

10.3.2 The Shape Option

This option allows the user to experiment with altering the default articulatory vocal tract model shape settings to reduce the error between the target formants and the model formants. This option is manually operated and is not the optimization process, which is the next step (phase). There are two purposes for this step in the software: to allow user experimentation to obtain an idea of the effect the articulatory model parameters have on the error and to speed up the optimization process. Sometimes an experienced user can select parameter settings faster than the optimization process, thereby reducing the optimization search time. Press the Shape button. A Shape Settings button appears, which in turn is to be pressed. The articulatory positions settings window shown in Figure 10.14 appears. The default values are shown for the nine articulatory model positions. The default values can be altered by the user using the sliders. The bottom of the figure shows the current frame and the total number of frames. These values correspond to the first frame and the three frames, respectively, for the marked be_seg_fmrk.mrk file shown in Figure 10.13. The begin time at the lower left is the beginning of the formant tracks, while the duration is the duration of the frame measured from the beginning time to the frame boundary. The Draw push button at the top, right allows the user to draw the articulatory vocal tract model for the position settings. An example for the current frame using the default position values appears in Figure 10.15, where the vocal tract model is drawn on the canvas for the articulatory speech synthesizer.

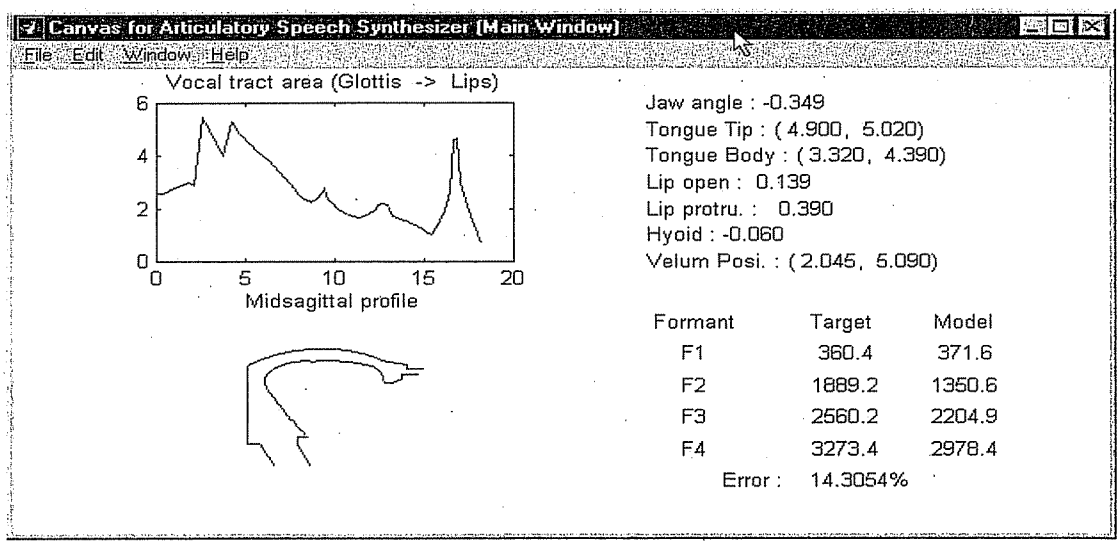


FIGURE 10.15 Articulatory vocal tract model.

If the user alters the default position values, then a new vocal tract model can be drawn by pressing the Draw button again. The Default button will restore the default values. The Clear button clears the canvas. It is not necessary to clear the canvas between successive draws, since the Draw command clears the canvas automatically. The Next button changes the frame to the next frame. The Previous button changes the frame back to the previous frame. The Sagittal button merely draws the number of sagittal sections on the vocal tract cross-sectional display. If the user alters the position values, then these values are not retained for the next frame, that is, the default values are restored automatically for the next frame. However, the user-set values are retained for the frame, unless altered again by the user.

Figure 10.15 shows the vocal tract area from the glottis to the lips, the nine position values for the articulatory model, the vocal tract cross-section (midsagittal profile), and the target and model formant values. The error for this configuration is shown as 14.3054%. This error can be easily reduced to less than 1% by adjusting the nine articulatory positions sliders.

The error is a weighted average of the absolute values of the difference between the target formants and the model formants. Define the first error term, e1, as the absolute value of the difference between the first target formant and the first model formant divided by the first target formant. Define similar error terms for the second, third, and fourth formants, e2, e3, e4. The error is (30)e1 + (30)e2 + (25)e3 + 15e4. Note that the first target formant cannot be zero.

10.3.3 The Optimization Option

The shape option must be selected and the window left open before the optimization option is selected. The default values for the shape option need not be altered and it is not required that the model be drawn. The software reads the values in the shape option window to start the optimization process. Pressing the Optimization button brings up two menu options: articulatory optimization setup and simulated annealing optimization. First, select the Setup Optimization button. This brings up the window shown in Figure 10.16.

The pull down menu offers four options for the dimension of the articulatory vector used in the simulated annealing optimization process (Figure 10.14): (1) eight position values for the tongue, lips, jaw, and hyoid, but with the velum closed; (2) nine position

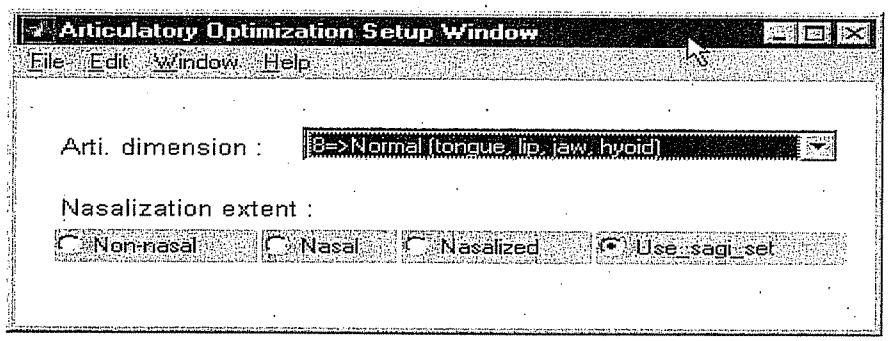


FIGURE 10.16 Articulatory optimization setup window.

values (the eight values above) and the velum, (3) eleven position values (the eight) and the pharynx (three values); and (4) twelve position values (the eight) and the velum and the pharynx (see Appendix 11 for more details). Briefly, as mentioned in Appendix 11-D, the eight position vector is best for front vowels; the nine position vector is best for nasalized front vowels; the eleven position vector is best for middle, back, and semivowels; and the twelve position vector is best for nasalized vowels. Usually, the velum position is set at a default position for nasal, non-nasal, or nasalized phonemes, but it can be optimized for some phonemes. The dimensions of the lower pharynx can also be optimized. These dimensions appear in Figure A11.4, and are given as the anterior–posterior movements of K and H (glk and wh in the software) and the height difference between K and H (hkl). When the pharynx is included, these three parameters, wh, glk, and hkl, are optimized by the software. There are default values for these parameters within the software.

The Nasalization Extent buttons allow the user to set the velum position for nasalization. The use_sag_set refers to the velum angle slider in Figure 10.14. This setting uses the value of the velum angle set by the user in that figure. The non-nasal setting closes the velum; that is, the slider setting is zero. The nasal setting is for a relatively large velum angle, namely, a slider setting of 492. The nasalized setting is for a moderate velum angle, namely, a slider setting of 324. The nasal tract design is included in the software, as discussed in Appendix 11. Appendix 11 shows some examples of the effect of no nasal tract, nasal tract coupling, and both nasal tract and sinus coupling; a sinus model is included.

After the user selects the desired articulatory vector dimension and nasalization, then select the Simulated Annealing Optimization button under optimization in the main window. This brings up the optimization window shown in Figure 10.17.

The annealing parameters that control the simulated annealing algorithm include the initial temperature T, the temperature reduction coefficient r_T, the number of steps to adjust the step length vector N_S, the number of step adjustments at each temperature N_T, the number of successive temperature reductions to test for termination N_ε, a small constant used for the termination criterion η, and the maximum number of function evaluations N_{tot}. The analogues between the annealing process and the articulatory problem can be identified as follows. First, the percentage of the weighted least-absolute-value (l_1-norm) error distance, equation (A11.3.3.1.3), corresponds to the energy of the material. The articulatory vector, equation (A11.3.3.1.1), corresponds to the configuration of particles. The change of articulatory parameters corresponds to the rearrangement of particles. Finding a near-optimal articulatory vector corresponds to finding a low-energy configuration. The temperature of the annealing process, T, becomes the control parameter for the speech inverse filtering process. Second, the Metropolis algorithm corresponds to the random fluctuations in energy. Third, the temperature reduction coefficient r_T corresponds to the cooling rate. Fourth, the finite number of moves at each downward control temperature value, $(N_S)(N_T)$, corresponds

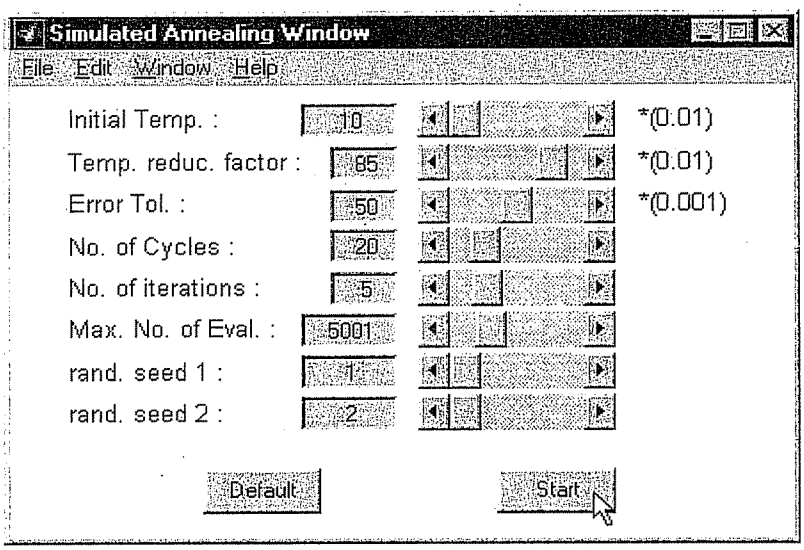
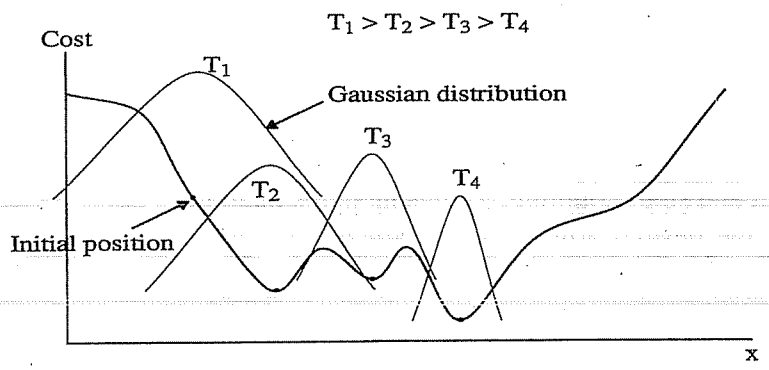


FIGURE 10.17 Simulated annealing window.

to the amount of time spent at each temperature. Reasonable values of the parameters (Table A11.4) are used as defaults for the optimization process, which are shown in Figure 10.17. However, a guideline of the optimization process is given in Appendix A11-D along with two examples. See Appendix 11 for more details. Figure 10.18 provides a summary of the effect of the temperature reduction factor on the simulated annealing algorithm and the Metropolis algorithm.



T: Control parameter (artificial temperature)

x: Articulatory parameter, e.g., tongue tip x coordinate

Metropolis Algorithm:

$$p(\Delta E) = 1.0, \qquad \text{for } \Delta E \leq 0,$$
$$= e^{\left(\frac{-\Delta E}{T}\right)}, \qquad \text{for } \Delta E > 0.$$

$$\Delta E = E(\text{new state}) - E(\text{current state})$$

FIGURE 10.18 The temperature reduction factor and the simulated annealing algorithm.

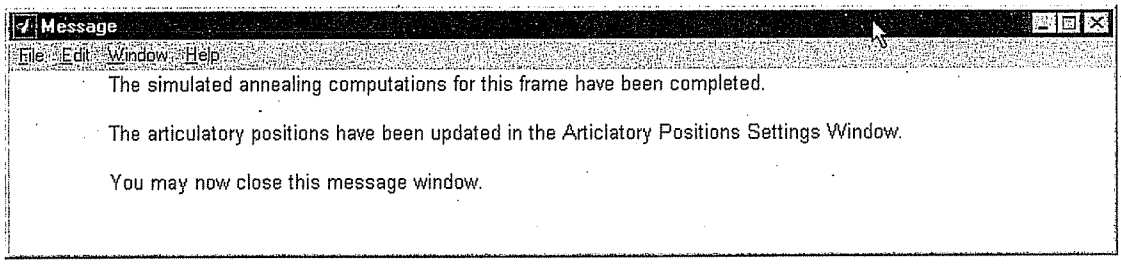


FIGURE 10.19 Message window stating that the frame calculations have been completed.

The parameter values shown in Figure 10.17 achieve a error on the order of 0.02%. But the optimization process takes on the order of 45 minutes with a 300 MHz machine. The parameter values can be altered by the user. For example, the optimization process is greatly speeded up by changing the number of cycles to 10 and the maximum number of iterations to 100. This results in an optimization on the order of a few minutes. However, such a change does not achieve as small an error as the default values because the search procedure is terminated prematurely. However, the error is still small, being on the order of 1% to 5% or so, compared to 0.02% or so.

To start the optimization, press Start in Figure 10.17. Figure 10.15 appears, where the error is shown as 14.3054%. The simulated annealing algorithm commences, and a succession of figures similar to Figure 10.15 are drawn on the canvas. Each successive figure has a smaller error. At the termination of the optimization process, a message window appears as shown in Figure 10.19.

At the same time Figure 10.20 appears, which is the final optimized articulatory configuration for the first target frame. The error is 0.0270%. The target and model formants are nearly identical in value. The vocal tract area is drawn along with the midsagittal profile and the values of the articulatory positions are given. The articulatory position settings in Figure 10.14 are also updated automatically and shown in Figure 10.21.

The user now presses the Next button in Figure 10.21 to advance the frame to frame 2 for optimization. The final articulatory position settings for frame 1, shown in Figure 10.21,

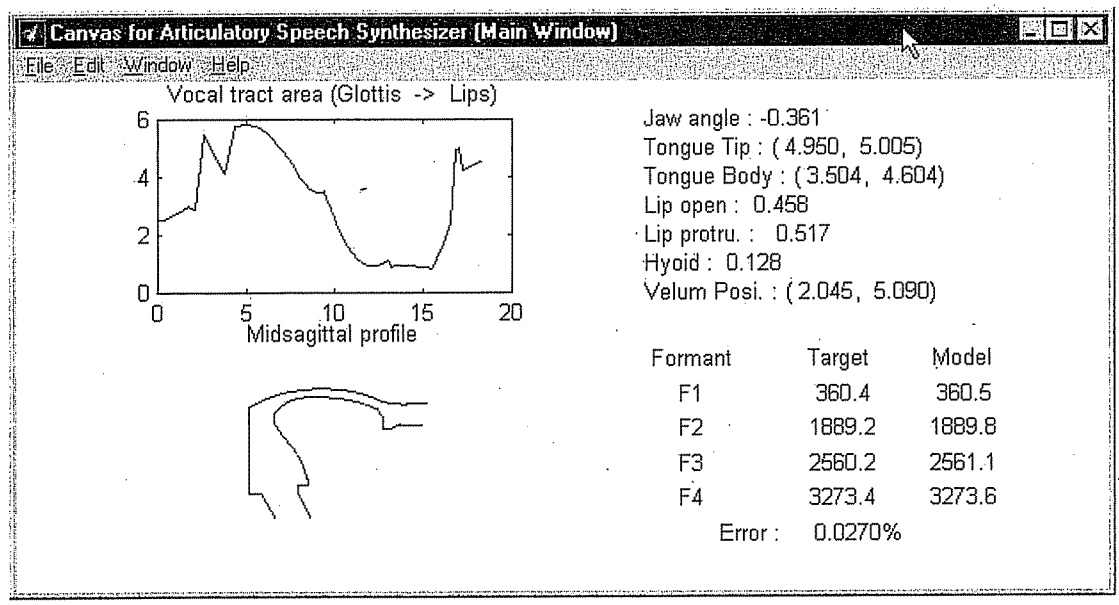


FIGURE 10.20 Final articulatory vocal tract model for frame 1.

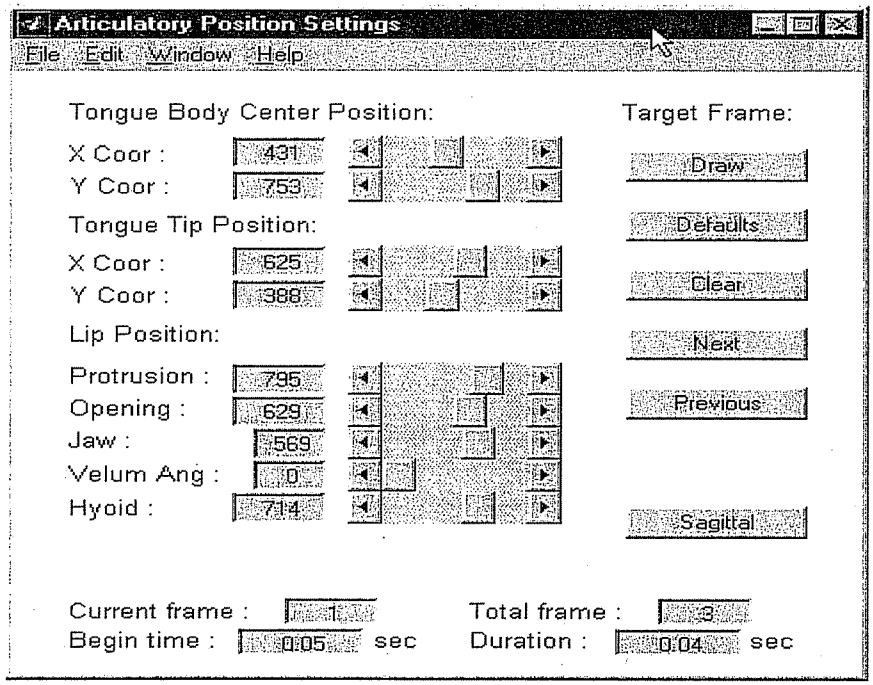


FIGURE 10.21 Final articulatory position settings for frame 1 after optimization.

are retained for frame 2. The reason for retaining the position settings from the last frame is that presumably, the position settings for the next frame are more similar to those of the last frame than the default values. This should speed-up the optimization process. The user can alter the simulated annealing algorithm values in Figure 10.17 or retain the default values. Press Start and the optimization process starts for frame 2. This time, the error begins at 1.3% and continues to reduce until the final error is 0.0385%. The same procedure is followed for frame 3. The final error is 0.0237%.

Upon completion of the optimization of the third and final frame, two message windows appear. One is the same as that shown in Figure 10.19. The other is shown in Figure 10.22.

The user now presses the File button in the main window and selects the Save Articulatory Parameters button to save the optimized articulatory parameter vector to a *.art file in the data folder. The art designation is for articulatory parameter vector. In this case, the saved file is called be_seg_opt.art. (The articulatory parameters vector is retained in working memory even if the file is not saved.) An *.art file can be loaded as an articulatory parameter vector, instead of a formant track file. In such a case, the Shape and

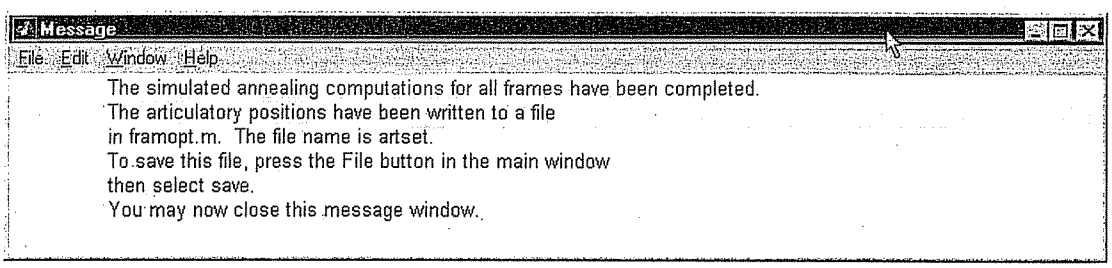


FIGURE 10.22 Message window stating that all simulated annealing computations are completed.

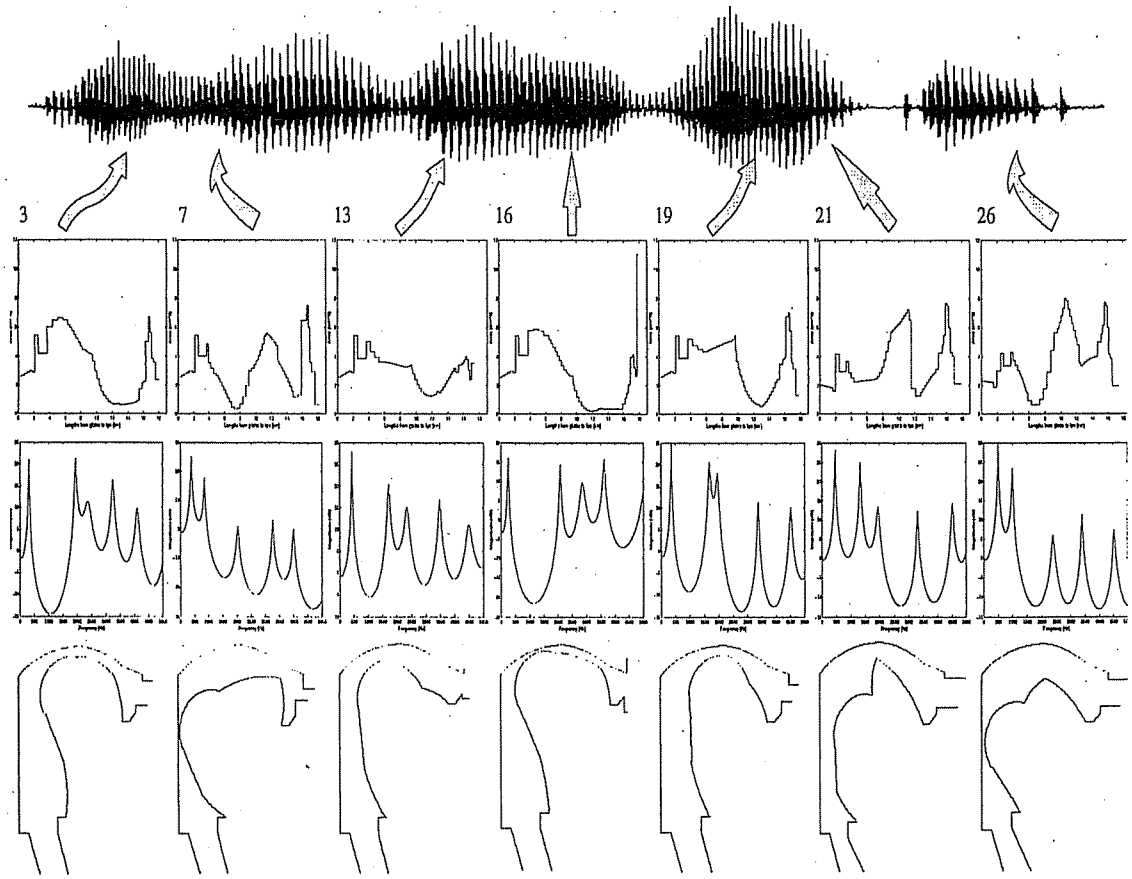


FIGURE 10.23 A summary of speech inverse filtering for the sentence, "We were away a year ago."

Optimization buttons are turned off (and therefore skipped) and the user progresses directly to the excitation option, which is described as follows.

Figure 10.23 is a summary of the speech inverse filtering results for the sentence, "We were away a year ago," using 26 frame boundaries. This example also appears in Appendix 11-D. The top panel in Figure 10.23 shows the time waveform for the sentence, the next panel down shows the vocal tract area functions, the next panel presents the formants, and the bottom panel shows the vocal tract cross-sectional (midsagittal) shapes for each frame. The results for only a few selected frames are presented due to space limitations. However, this figure is an illustration of how the toolbox is used to obtain the vocal tract configurations by minimizing the error between the target and model formants for each frame.

10.3.4 The Excitation Option

The excitation option can be started at any time. Neither a formant track file nor an articulatory parameter file needs to be loaded. However, the usual procedure is to design the excitation waveform after the articulatory parameter vector has been obtained via optimization. Prior to starting the excitation option, the user can close any open windows, except the main menu window and the canvas window. Next, press the Excitation button and the LF Model button. Figures 10.24, 10.25, and 10.26 appear.

In Figure 10.26, the user has pressed the Glottal Excitation Place button, which in turn activates the excitation mode options, namely, highlighting the Voiced, Unvoiced,

FIGURE 10.24 Excitation source menu window.

and Mixed Excitation buttons. If the Glottal Excitation button is activated, then the Vocal Tract (VT) button cannot be activated. This latter option is explained later. After the Glottal Excitation button is activated (pressed, i.e., the box is filled), the user can now select either voiced, unvoiced, or mixed excitation mode. These options are explained now.

10.3.4.1 Voiced Excitation In the following, we illustrate the construction of a voiced excitation waveform. The user presses the Voiced button in Figure 10.26, which activates (highlights) the Jitter and Shimmer button, the Aspiration Noise button, and the Subglottal Model button. Imagine that the user has pressed these three buttons as well, as shown in Figure 10.27. Note that once any of the buttons is pressed (activated), they can be cleared (deactivated) by pressing them again. Also note that at this time, there is only one type of excitation model waveform, namely, the LF model.

Pressing the Voiced button brings up the menu window in Figure 10.28, which allows the user to adjust the LF waveform model parameters as well as the fundamental frequency of voicing, the gain, and another gain parameter, g0. If the user changes F0, then T0 changes automatically, and vice versa. The LF model is described in Appendix 7. The gain (av) determines the gain for the integrated LF model waveform. The parameter g0 provides additional flexibility to allow the user to scale the volume velocity waveform when mixed with aspiration noise. The default values are shown in Figure 10.28.

The jitter and shimmer menu window appears in Figure 10.29. The jitter and shimmer sliders specify the maximum jitter and shimmer values as a tenth percent of the fundamental frequency of voicing, F0, and the gain (av), respectively. The jitter and shimmer settings are only valid for sustained voicing. Jitter and shimmer are deactivated for the unvoiced option. The jitter and shimmer filter sliders specify the filter coefficients for the psuedo-random number generators, respectively. A negative coefficient means a high-pass filter, a zero coefficient represents no filtering, and a positive coefficient is a low-pass filter. The absolute value of the coefficient must be less than one.

Aspiration noise can be added to the excitation waveform through the use of the parameter window shown in Figure 10.30. The gain slider allows the user to set the gain for the noise, while the filter slider sets the filter coefficient for the aspiration filter in a similar manner to that described for the jitter and shimmer option.

Figure 10.31 summarizes the voiced excitation model.

The subglottal model is available only for voiced excitation, and is shown in Figure 10.32. There are two aspects to the subglottal model: the upper portion and the lower

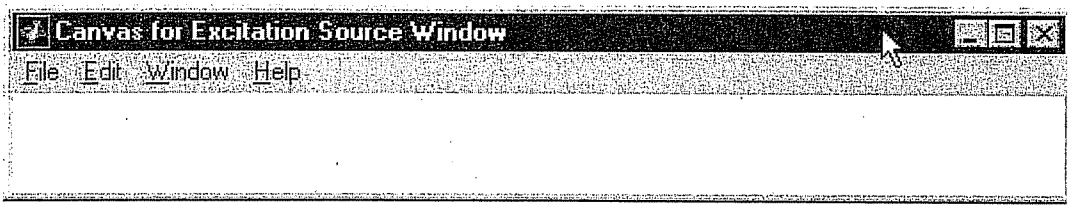


FIGURE 10.25 Canvas for the excitation source window.

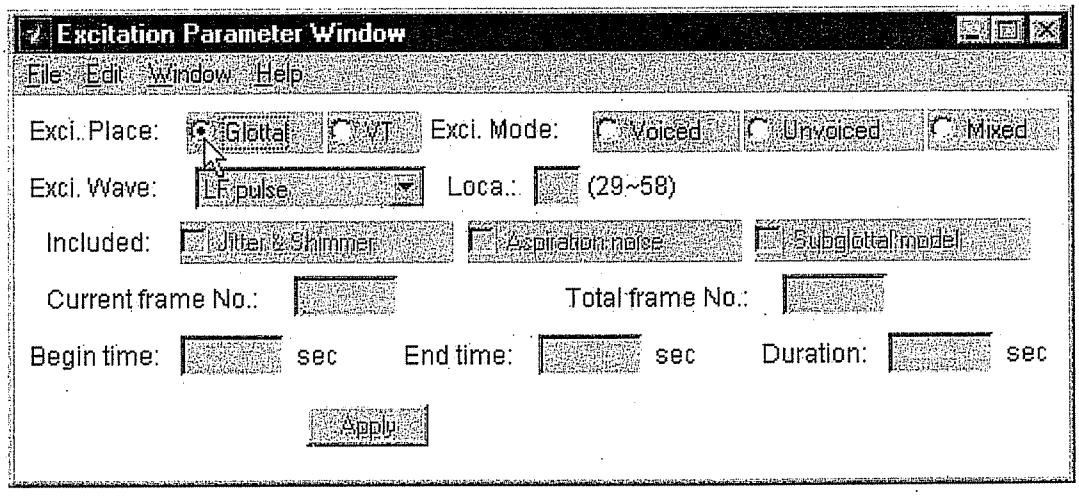


FIGURE 10.26 The excitation parameter window.

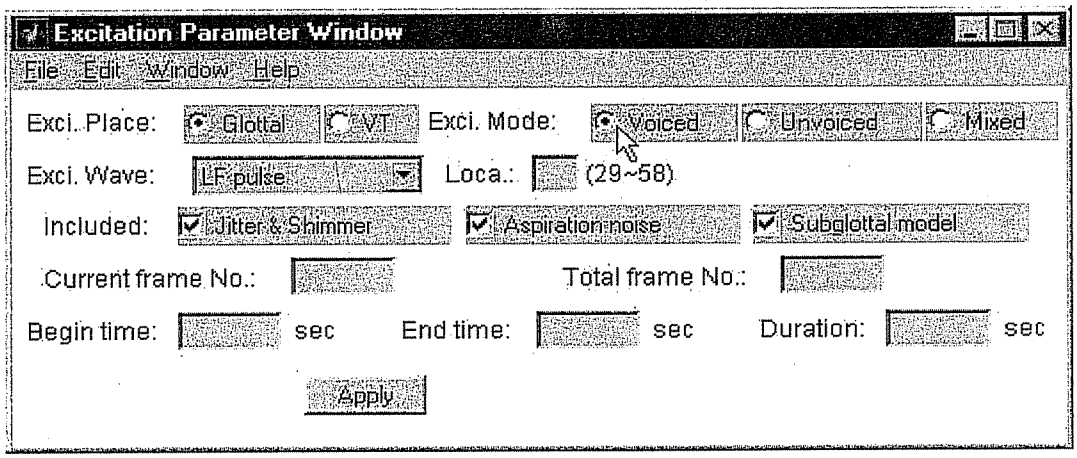


FIGURE 10.27 Excitation parameter window for voiced excitation with jitter and shimmer, aspiration noise, and subglottal model.

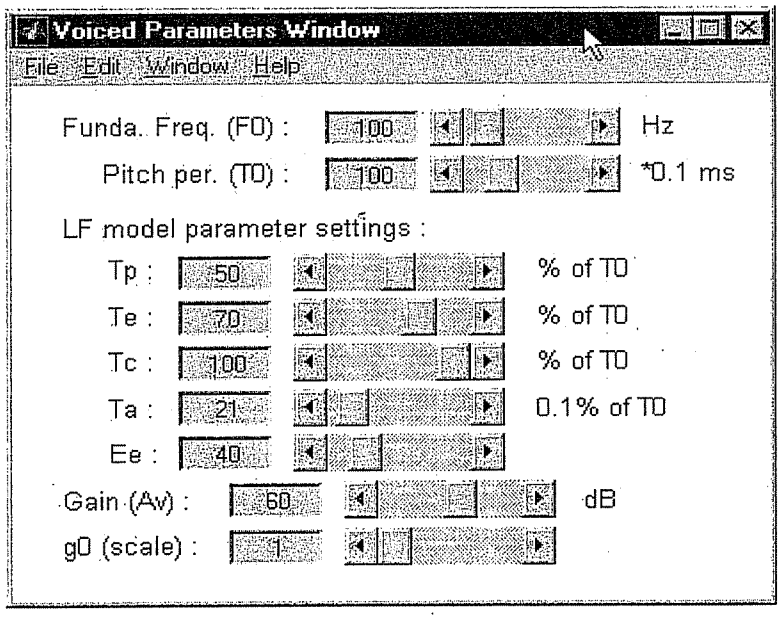


FIGURE 10.28 The voiced parameters window.

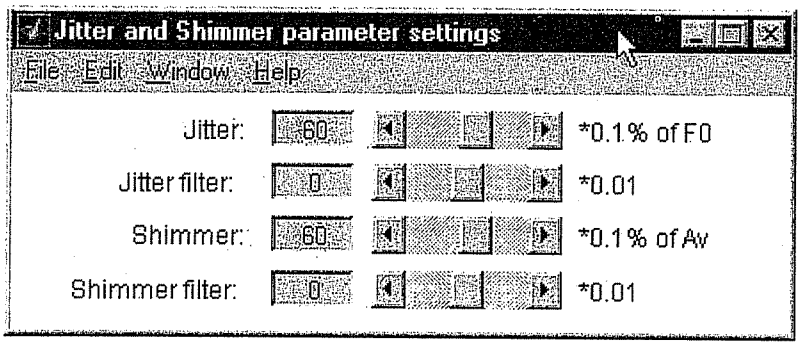


FIGURE 10.29 The jitter and shimmer parameters window.

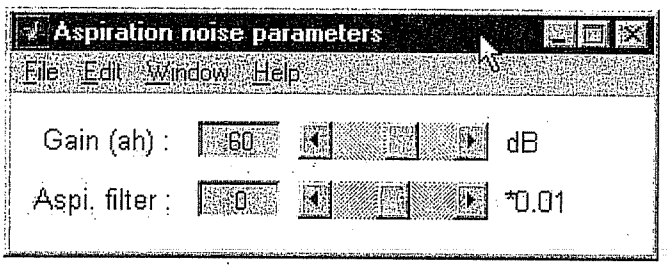


FIGURE 10.30 Aspiration noise parameters window.

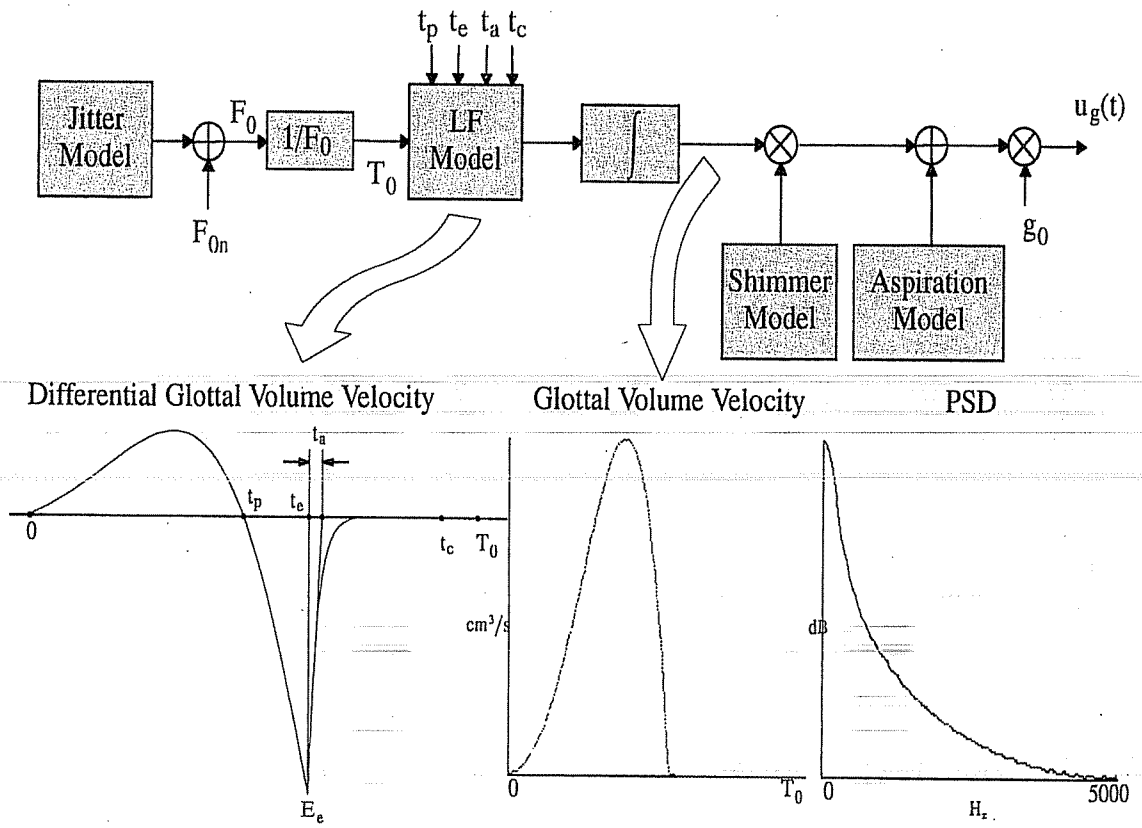


FIGURE 10.31 The voiced excitation model.

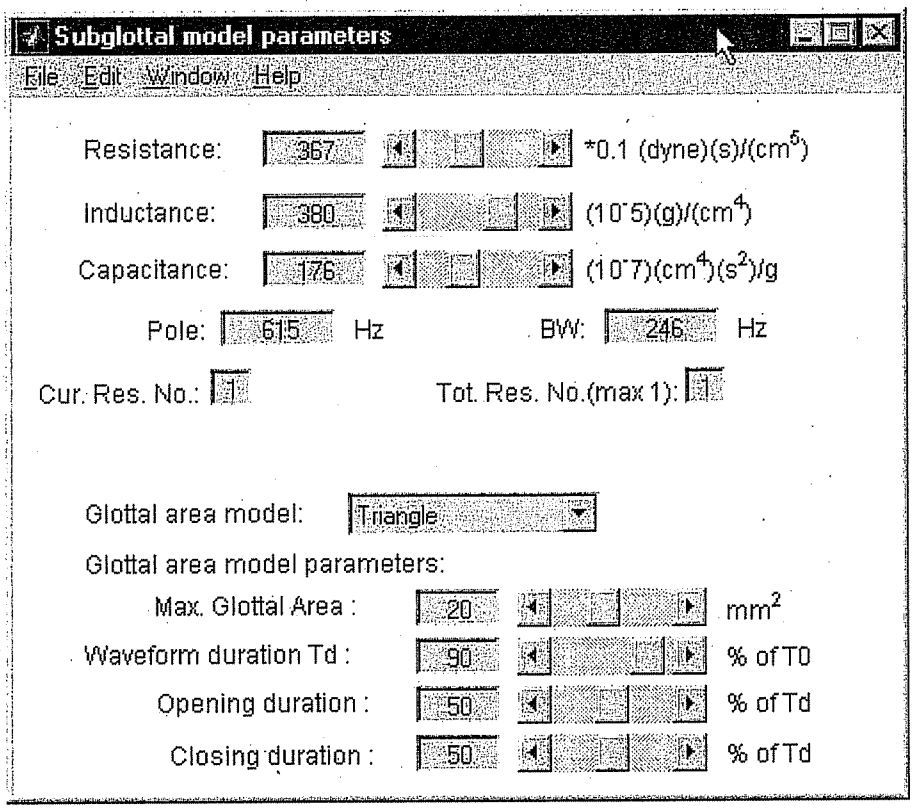


FIGURE 10.32 The subglottal parameters window.

portion. Since Foster RLC circuits are used to model the subglottal tract, the upper portion of Figure 10.32 enables the user to specify the RLC values and calculates and displays the resonant frequency (pole) and bandwidth. The current resonator is entered by the user and there is only one resonator allowed by the model at this time. For voiced and mixed excitation, the glottal area (lower) is a time-varying function. The user can select one of three glottal area functions using the pull-down window: triangular, sinusoidal, or raised cosine, as shown in Figure 10.33. For each glottal area model, the maximum area (in mm^2) and the waveform duration (in percent of T0) is to be specified by adjusting the appropriate sliders. For the triangular and raised cosine glottal area models the waveform can be unsymmetrical. This can be accomplished by adjusting the opening duration and closing duration sliders. Adjusting one slider automatically adjusts the other. For the sinusoidal model the waveform is always symmetric, so these sliders have no effect. For the unvoiced option, the glottal area is a constant. So only the maximum glottal area slider is activated to allow the user to specify the glottal opening area.

Once the user has set the desired parameter values, then type in the current frame number, the total number of frames, the beginning time, the duration, and the ending time (beginning time plus duration) as shown in Figure 10.34. The beginning time and duration are available from Figure 10.21 and subsequent such figures for each frame. Then press the Apply button. The excitation waveform can be viewed by pressing the Draw button in Figure 10.24, the main menu excitation source window. For the parameter settings used here, the excitation waveform appears as shown in Figure 10.35, which is the canvas for the excitation source waveform.

Figure 10.35 shows the LF model waveform, which is the differentiated glottal volume velocity waveform, the glottal waveform (glottal volume velocity), the glottal area model waveform, and the power spectral density (PSD) for the excitation waveform.

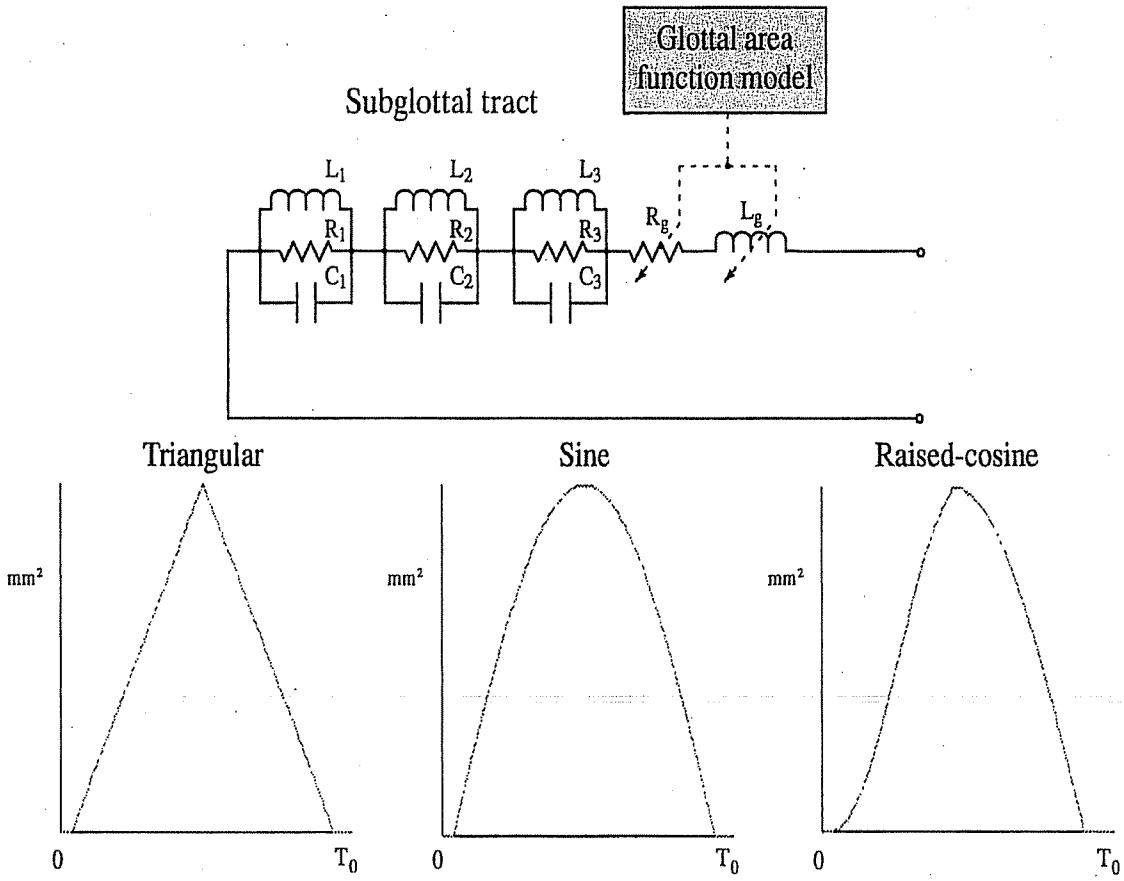


FIGURE 10.33 Subglottal system and glottal impedance model.

The user can design additional excitation waveforms for other frames, but one excitation waveform is sufficient for synthesis and is the recommended procedure. The jitter and shimmer are included in the synthesis process as well as the aspiration noise and the subglottal model. The excitation waveform can be saved by pressing the File button in Figure 10.24 and selecting the Save option. Excitation waveforms are saved as *.src files,

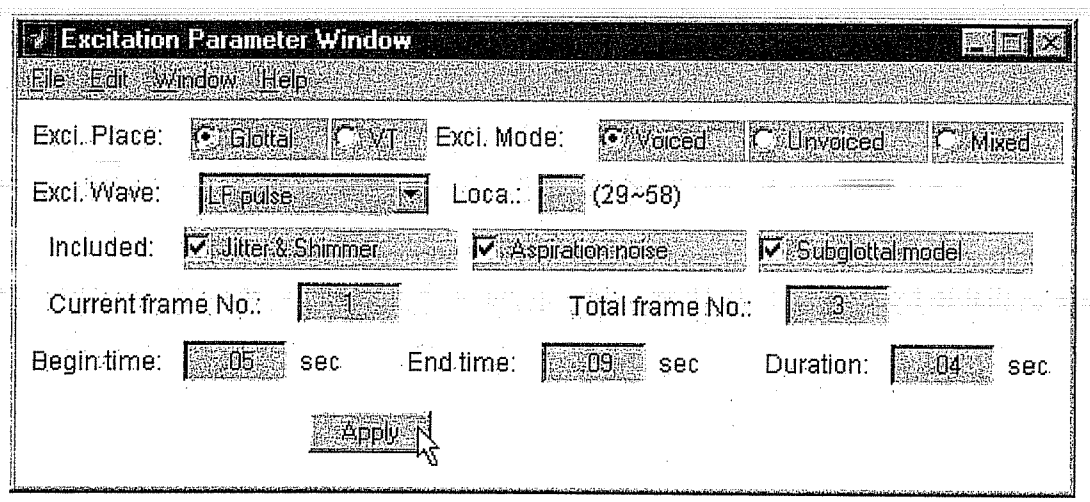


FIGURE 10.34 Excitation parameter window with the frame and time parameters set.

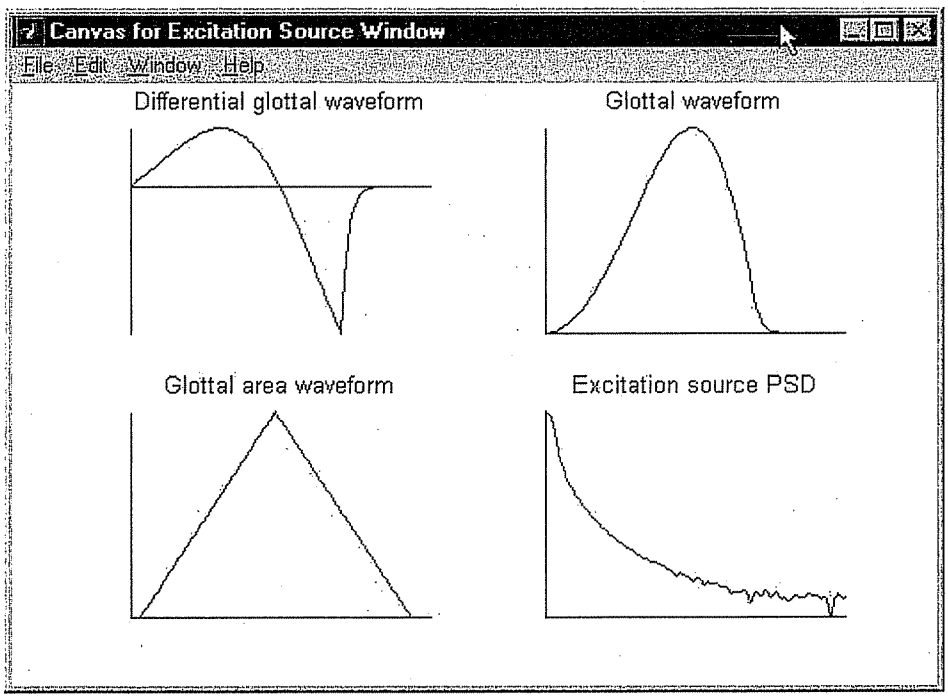


FIGURE 10.35 Excitation waveform.

where the * is replaced by a name selected by the user. A saved excitation waveform can be loaded by pressing the File button and selecting the Load option. Once a saved excitation file is loaded, the windows that were used to design the excitation open automatically. For example, if the excitation waveform was designed as voiced with jitter and shimmer and aspiration noise, then these windows also open, showing the design parameter values. Thus, the parameters of a loaded excitation waveform can be altered and saved as another file if desired. The Next and Previous buttons in Figure 10.24 are to allow the user to examine the parameter values and plot the excitation waveforms of a designed excitation or a loaded excitation waveform. These two options are typically not used since usually only one excitation waveform is used in the synthesis process. The Clear button erases the canvas window. An excitation waveform must be designed or loaded before the synthesis process is started. Similarly, an articulatory parameter vector must be designed or loaded before the synthesis process is initiated.

10.3.4.2 Unvoiced Excitation The options for unvoiced excitation are similar to that for voiced. However, certain options are not available, while others are. Pressing the Unvoiced option button in Figure 10.26 activates only the aspiration noise and subglottal model options. The jitter and shimmer option is not available for unvoiced excitation. The unvoiced option opens the turbulence noise parameters window shown in Figure 10.36. This window allows the user to design the turbulence excitation and set the place of the turbulence at 1) downstream from the maximal constriction point (the default); 2) the center of the maximal constriction; 3) upstream from the maximal constriction; or 4) distributed along the constriction region. See Appendix 11 for more details. The turbulence gain is adjustable, as is the critical Reynolds number (which sets the threshold at which the volume velocity becomes turbulent flow) and the magnitude of the glottal volume velocity, which sets the volume velocity at a constant value for the unvoiced excitation. If the aspiration and subglottal options are selected, then these are set as described for voiced excitation.

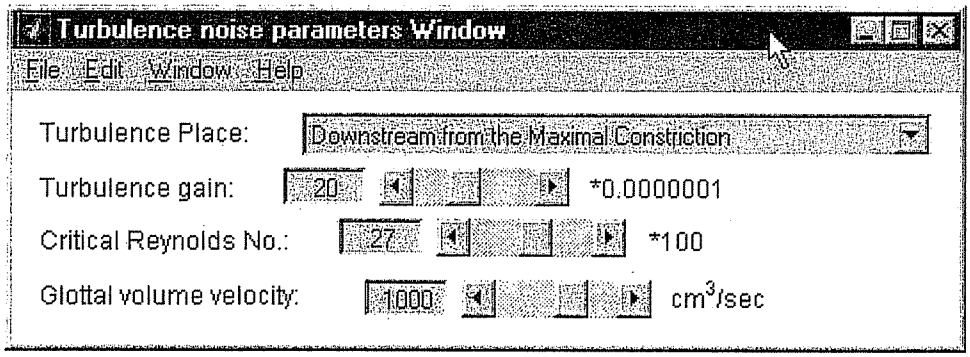


FIGURE 10.36 Turbulence noise parameters window for unvoiced excitation.

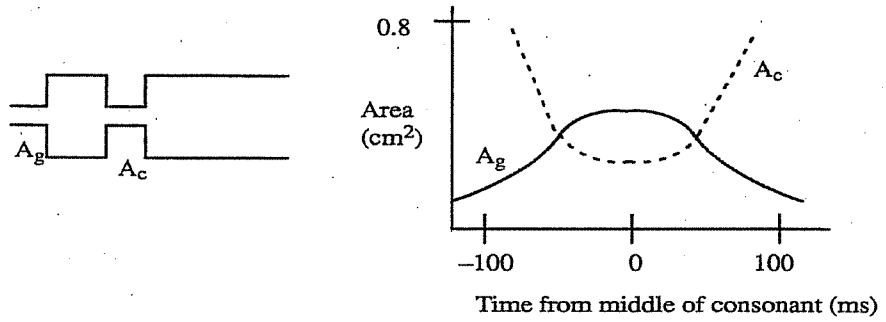
The unvoiced excitation can be saved or loaded as described previously. The waveform can be drawn by pressing the Draw button in Figure 10.24, the excitation source window.

10.3.4.3 Mixed Excitation This option provides for both voiced and turbulence excitation. Pressing the Mixed Excitation button brings up both the voiced parameters window and the turbulence noise parameters window. The user can also select jitter and shimmer, aspiration noise, and the subglottal model, as for voiced excitation. The designed excitation waveform can be saved, loaded, and drawn as described previously.

10.3.4.4 Turbulence Excitation Models Appendix 11 discusses the development of a model for unvoiced, or turbulence, excitation. The elements of this model are presented in Figures 10.37 and 10.38. The model for the turbulence noise source can be located at the center of the constriction, immediately upstream or downstream from the constriction, or spatially distributed within the constriction.

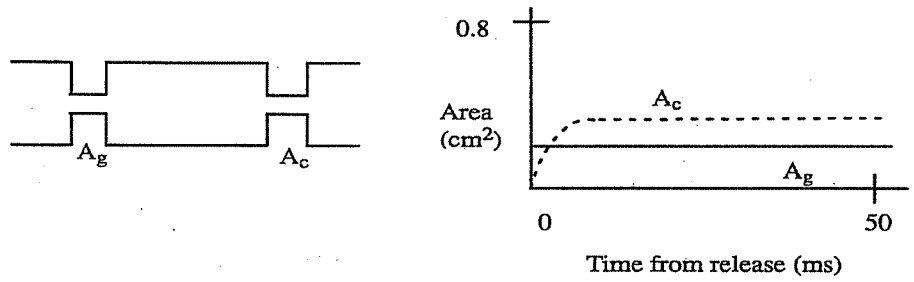
Examples of the vocal tract area and the transfer function for these unvoiced models are given in Appendix 11.

10.3.4.5 Excitation Within the Vocal Tract The final option available for the excitation waveform generation is the ability to use voiced excitation at a location within the vocal tract (see Appendix 11 for additional discussion of this option and its application to individuals who have no larynx to generate a voiced excitation). This is a special feature of this toolbox and will be useful to only a few researchers. The primary function of this option is to show that the excitation waveform can be designed (reshaped) to synthesize normal sounding speech even if the excitation is placed within the vocal tract, instead of at the glottis. To use this option select the VT button in Figure 10.26 instead of the Glottal Excitation button. This activates only the voiced option and the location option next to the excitation waveform on the second row in Figure 10.39. Press the Voiced option button. This opens the voiced parameters window and activates the jitter and shimmer and aspiration noise options as discussed previously for the voiced excitation option. The user may select either, both, or neither of these latter two options. The user must type in a vocal tract excitation location for this option. Figure 10.39 shows the location typed as 35. This designates that the voiced excitation is to occur at the 35th section of the vocal tract model, not at the glottis. The user designs the excitation waveform as described previously, enters the beginning time and other variables, and presses the Apply button. The designed waveform is plotted in Figure 10.40 by pressing the Draw button in Figure 10.24, the excitation source window. When the synthesis process is performed, the waveform excites the vocal tract at the 35th section and generates the appropriate synthetic speech with the excitation waveform at this vocal



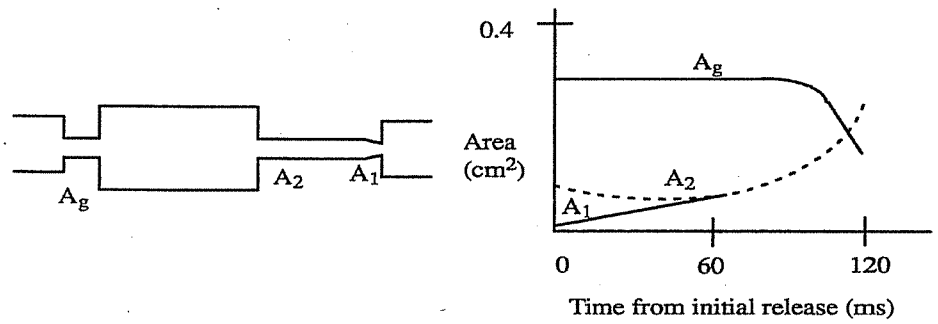
Fricative

(a)



Stop

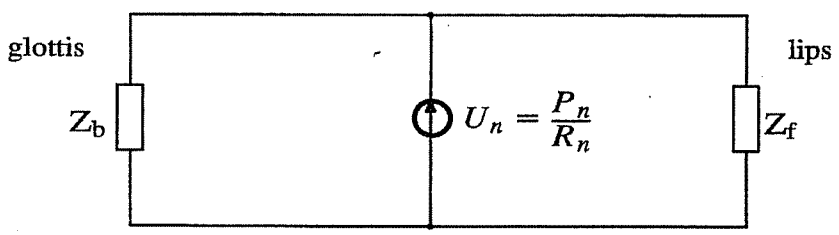
(b)



Affricate

(c)

FIGURE 10.37 Unvoiced (turbulence) excitation models for (*a*) fricatives, (*b*) stops, and (*c*) affricates.



$$R_n = \frac{\varrho |U_n|}{2A_c^2}, \qquad \text{Reynolds number } R_e = \frac{4\varrho^2}{\pi\mu^2} \cdot \frac{U_n^2}{A_c}$$

$$P_n = turbg \cdot rand \cdot \left(R_e^2 - R_{ec}^2\right), \quad \text{for } R_e > R_{ec}$$

$$= 0, \quad \text{for } R_e \leq R_{ec}$$

FIGURE 10.38 Turbulence noise source model.

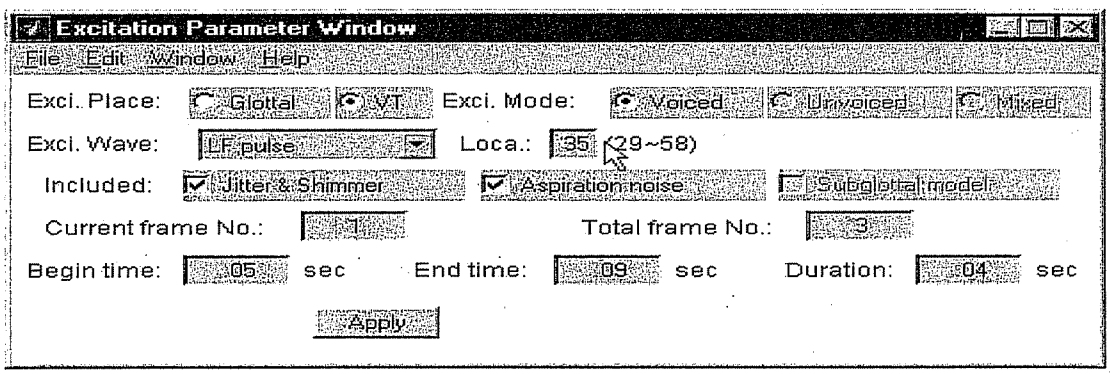


FIGURE 10.39 Excitation parameter window for vocal tract excitation at the 35th section.

tract location. Note that the designed excitation waveform provides the excitation; it is not prefiltered to compensate for the fact that the excitation is relocated.

While Appendix 11 discusses this feature of the articulatory speech synthesizer, we present several examples here. A model for calculating the transfer function with the excitation within the vocal tract is shown in Figure 10.41. The calculations for transfer function proceed from top to bottom in Figure 10.41. This model can be used to calculate a filter that can be used to prefilter the glottal excitation waveform to compensate for the fact that the excitation is relocated to a section within the vocal tract. The software does not provide the prefiltering; this must be done by the user. In the following examples if the excitation is relocated to a section within the vocal tract, it is not prefiltered unless otherwise stated.

Figure 10.42 shows the midsagittal profile for the vowel /AA/.

The optimized vocal tract cross-sectional area and the transfer function for the vowel /AA/ are shown in Figures 10.43 and 10.44.

Next, Figures 10.45 and 10.46 show the original and synthesized speech waveforms and spectrograms, respectively, with the excitation located at the 35th and 50th section of the vocal tract for the vowel /AA/.

Figure 10.47 shows the glottal pulse of the LF model immediately followed by the same glottal pulse prefiltered by the modified acoustic transfer function (top panel), the synthesized speech for the vowel /AA/ with the above excitation located at the 35th

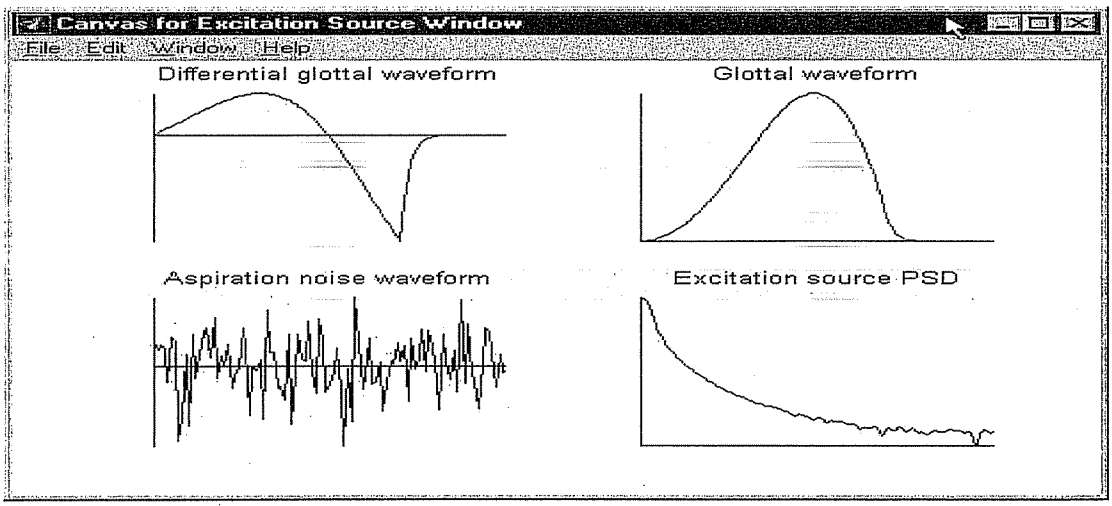
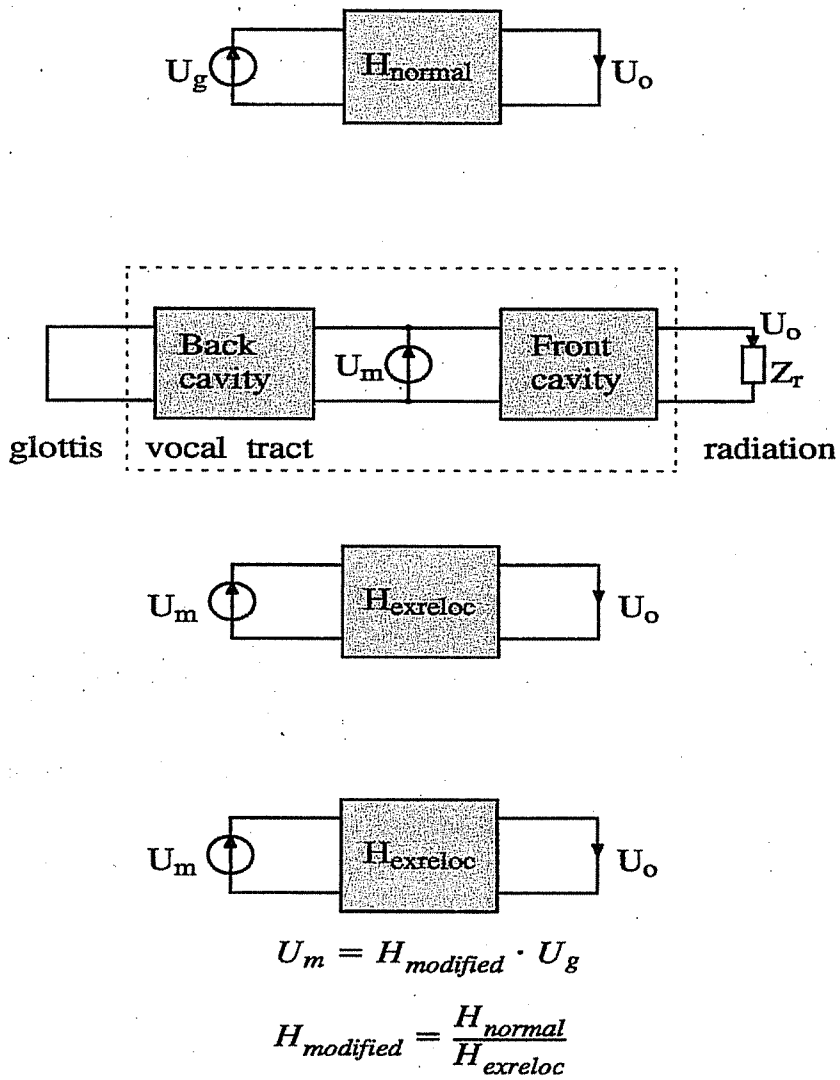


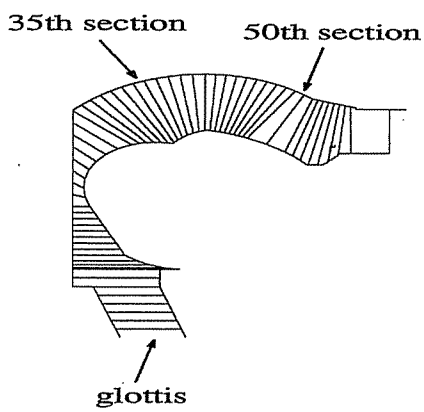
FIGURE 10.40 Excitation waveform for excitation at the 35th vocal tract section.



$$U_m = H_{modified} \cdot U_g$$

$$H_{modified} = \frac{H_{normal}}{H_{exreloc}}$$

FIGURE 10.41 Source model for excitation relocation in the vocal tract.



Midsagittal profile for vowel /a/.

FIGURE 10.42 Midsagittal profile for the vowel /AA/.

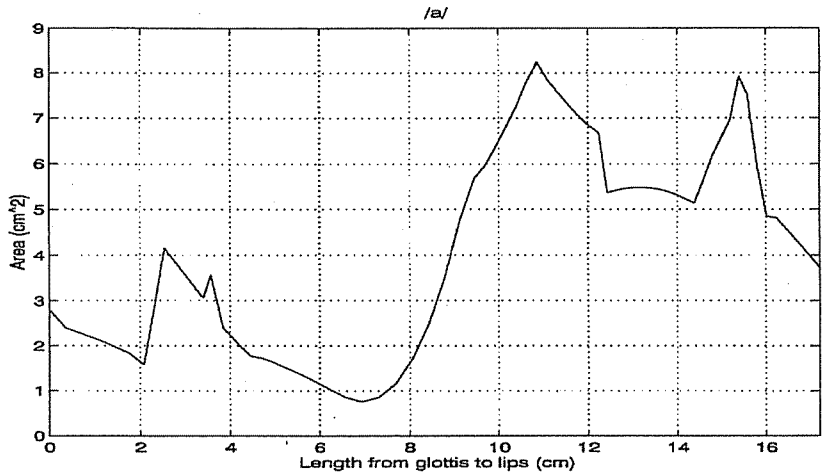


FIGURE 10.43 Optimized vocal tract cross-section area for vowel /AA/.

section of the vocal tract (middle panel), and the wideband spectrogram of the synthesized speech (lower panel). The purpose of this figure is to illustrate the effect of prefiltering the excitation waveform before exciting the vocal tract at the 35th section. In this case, the synthesized speech without prefiltering sounds very good, while the speech synthesized with the prefiltered excitation sounds like a tone. This is because the prefiltering introduces the fifth and sixth formants as artifacts. Consequently, the prefiltering needs to be improved.

The final two figures in this section, Figures 10.48 and 10.49, show the original and synthesized speech waveforms and spectrograms, respectively for the sentence, "We were away a year ago," with the excitation located at the 35th and 50th section of the vocal tract. The excitation is not prefiltered.

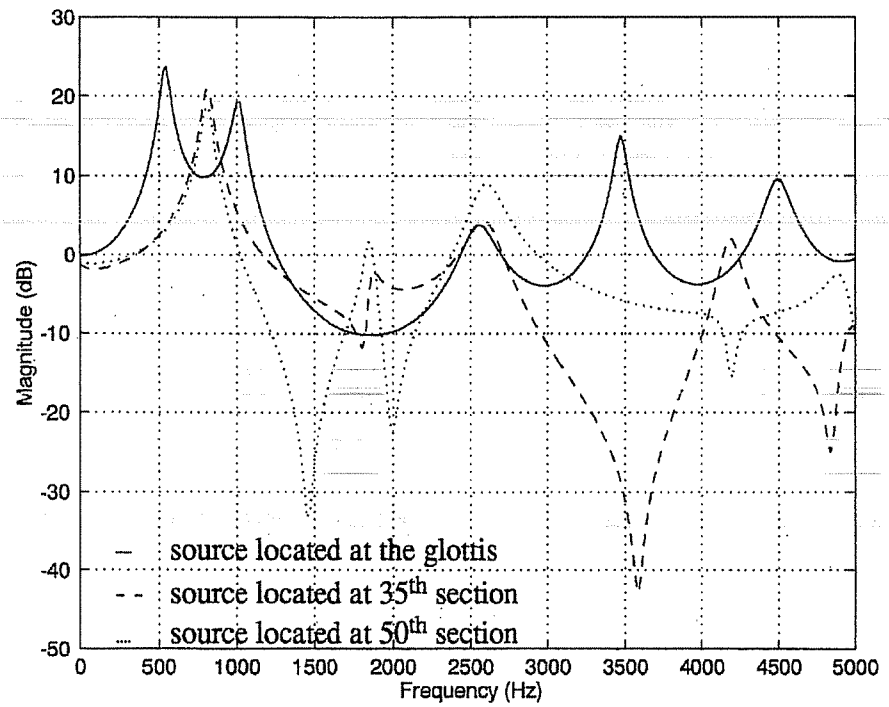


FIGURE 10.44 Transfer function for the vowel /AA/ with the source located at three locations.

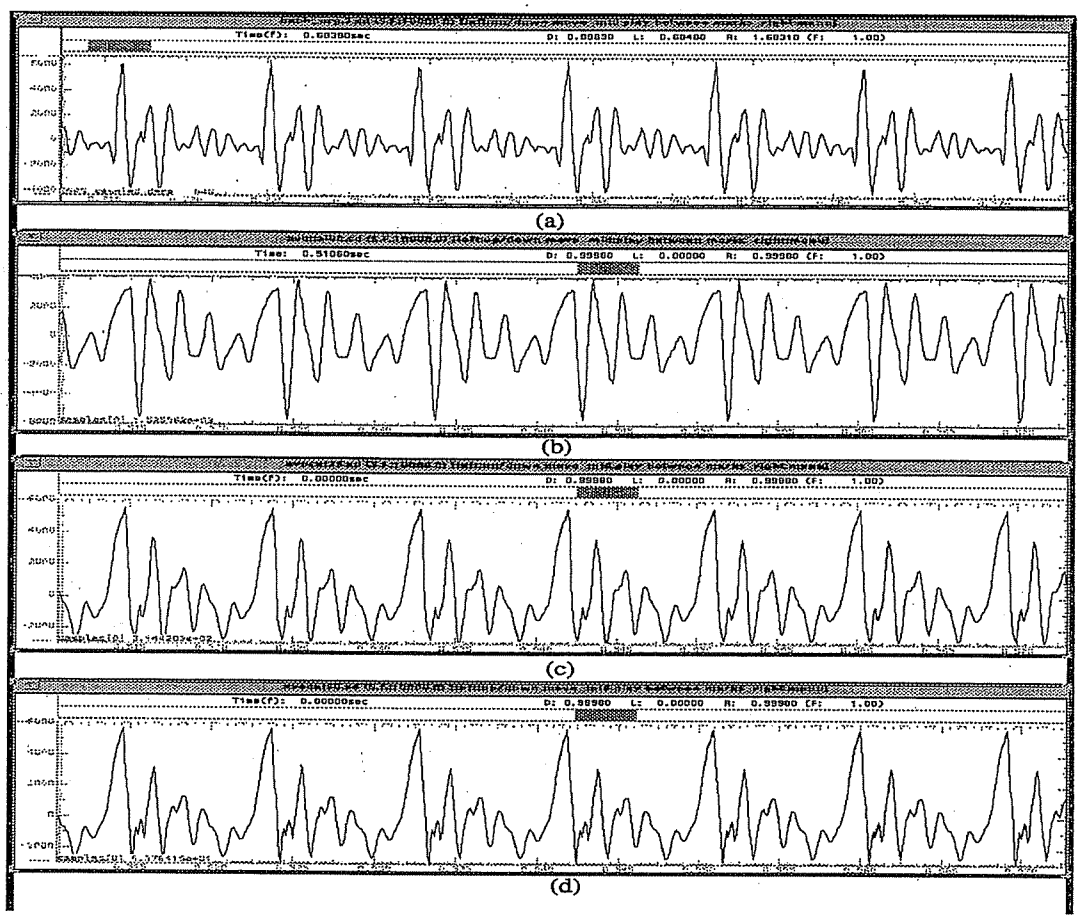


FIGURE 10.45 Original (*a*) and synthetic speech waveforms located with the source located at the,(*b*) glottis, (*c*) the 35th section, and (*d*) the 50th section of the vocal tract for the vowel /AA/.

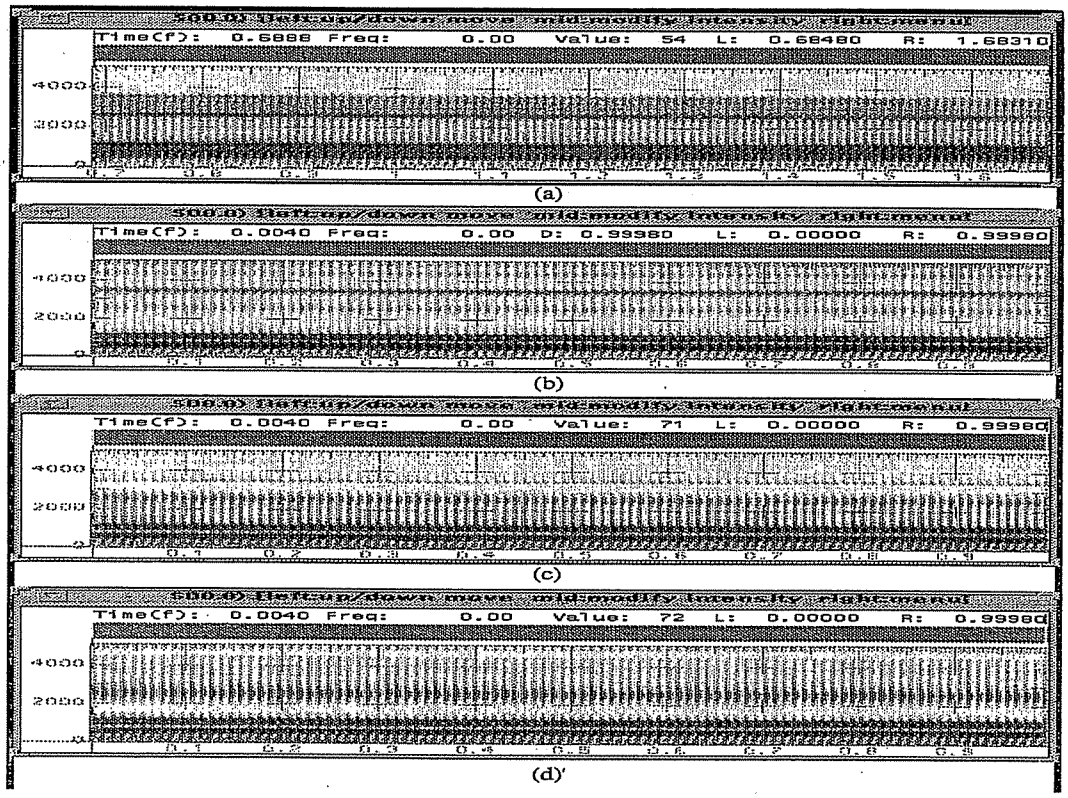


FIGURE 10.46 The spectrograms for the waveforms shown in Figure 10.45.

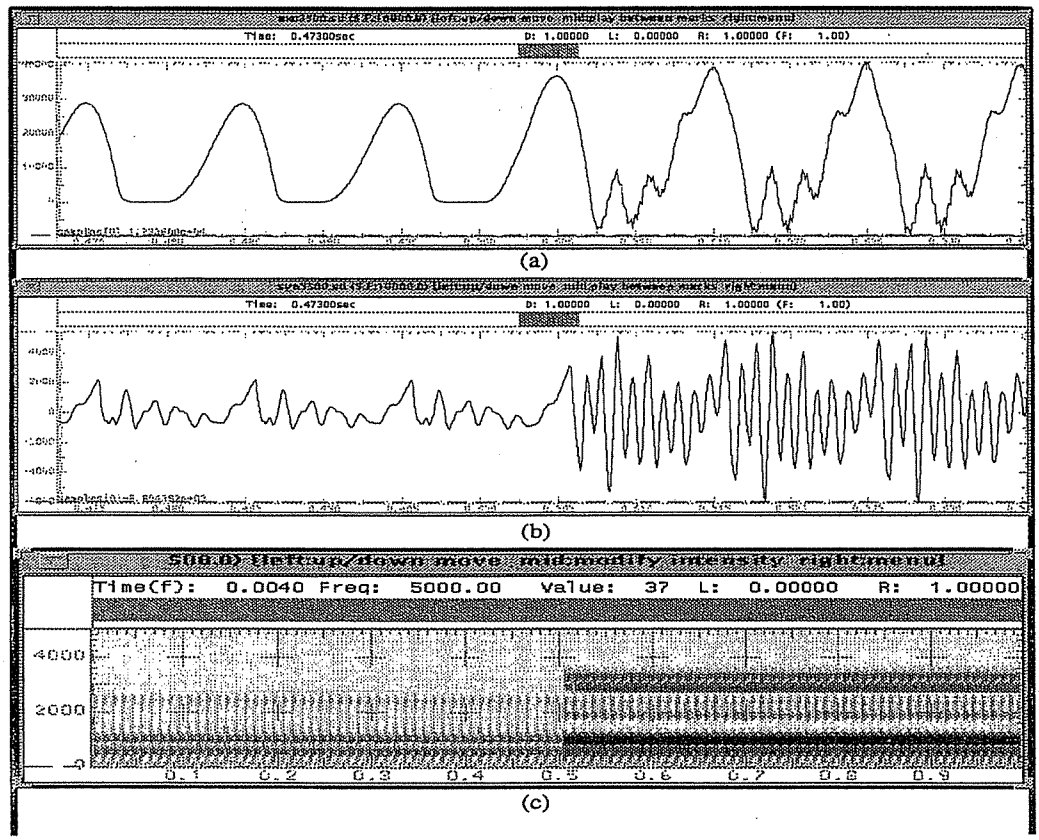


FIGURE 10.47 Illustration of prefiltering the excitation waveform before placing the excitation at the 35th section of the vocal tract for the vowel /AA/.

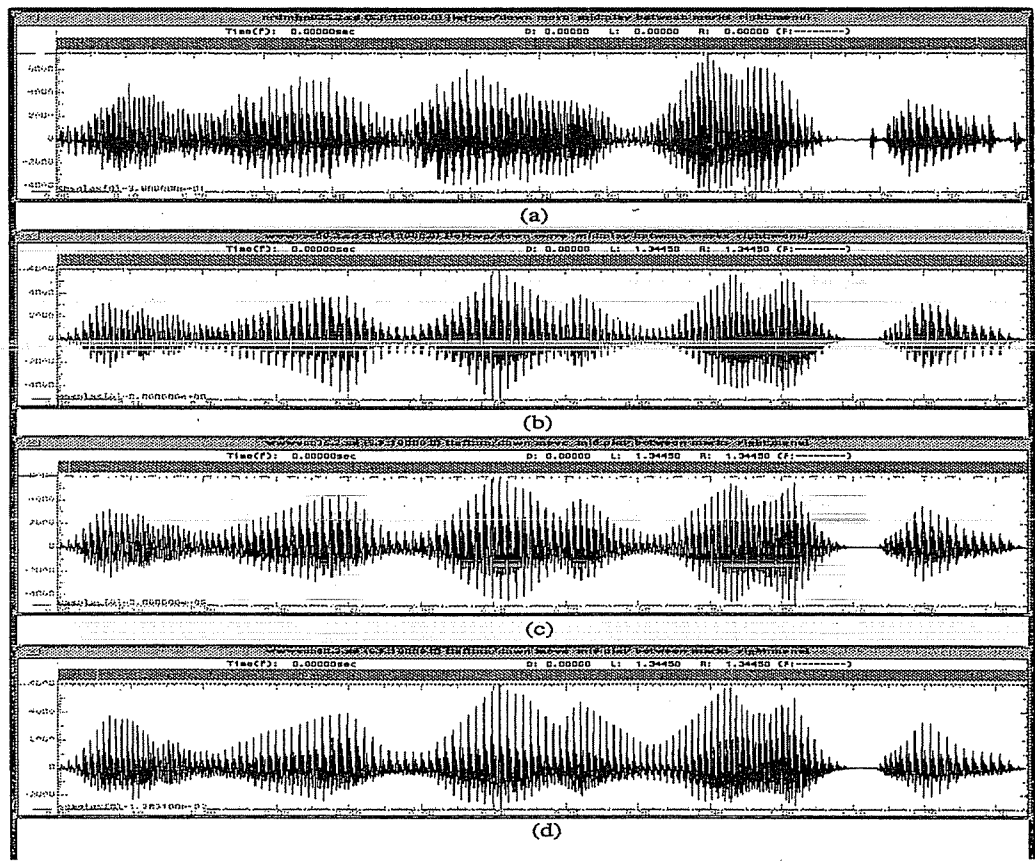


FIGURE 10.48 Original (*a*) and synthetic speech waveforms located with the source located at the, (*b*) glottis, (*c*) the 35th section, and (*d*) the 50th section of the vocal tract for the sentence, "We were away a year ago."

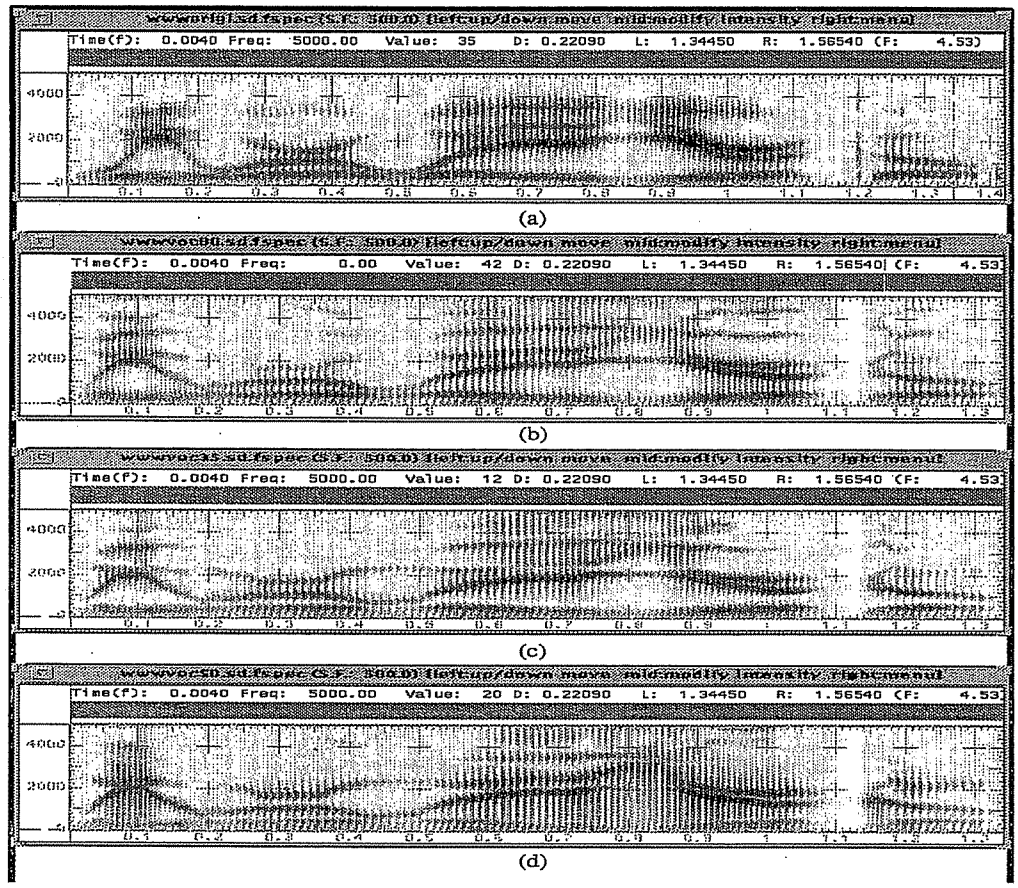


FIGURE 10.49 The spectrograms for the waveforms shown in Figure 10.48.

10.3.5 The Synthesis Option

10.3.5.1 Review of Steps Before Synthesis Before activating the synthesis option, several steps must be completed. The user must either load an articulatory parameter file obtained by following the optimization procedure, or load a target formant track file and mark the frame boundaries on the formant tracks. Next, the user opens the shape option, followed by the optimization option, thereby obtaining an articulatory parameter vector. Then the user either loads a previously designed excitation waveform or designs a new excitation. In summary, the working memory must contain 1) an articulatory parameter vector; and 2) an excitation waveform prior to the initiation of the synthesis process.

10.3.5.2 Starting the Synthesis Option Pressing the Synthesis button in the articulatory speech synthesizer (main window), closes all windows opened during the excitation option, and opens the three windows shown in Figures 10.50, 10.51, and 10.52, which are the articulatory synthesis menu window, the canvas for the articulatory synthesis, and the articulatory synthesis parameters window.

The menu selections in the articulatory synthesis window are start, animate, play, print, clear, and cancel. The Cancel button cancels the synthesis option and erases the

FIGURE 10.50 The articulatory synthesis menu window.

FIGURE 10.51 The canvas for the articulatory synthesis menu window.

three windows shown in Figures 10.50, 10.51, and 10.52. The Animate button opens the window shown in Figure 10.53 and plays a movie of the midsagittal profiles of the articulatory configuration for the articulatory parameter file in memory, which in this case, is the be_seg_opt.art file. For this case, the movie contains only three frames. The movie is repeated three times, ending with the last frame. The Play, Print, Clear, and Save buttons are explained later.

Prior to pressing the start button in Figure 10.50, the user selects one of three display options in Figure 10.52 and the synthesis sampling frequency. The default settings are shown as display choice 1 and a sampling frequency of 20 kHz. The three display options are:

1. The vocal tract cross-sectional area function, the midsagittal vocal tract outline of the current target frame, the target and model formants, the error, and the synthetic speech waveform.
2. The articulatory trajectories and the synthetic speech waveform.
3. The target and excitation frame data; the acoustic transfer function, the excitation waveform and its power spectrum; the pressure and volume velocity waveforms at the 20th, 30th, 40th, and 50th sections in the vocal tract; and the synthetic speech waveform.

After the Start button is pressed, it takes approximately 30 seconds before the display appears. The interpolation of the articulatory parameters is linear as explained in Appendix 11. The synthesis sampling rate is selected as 10, 20, 30, 40, 50, or 60 kHz. There are a total of 60 sections in the articulatory vocal tract model. This number is fixed and cannot be altered by the user. For display option 3, the pressure, volume velocity, and speech waveforms are updated (refreshed) every 400 samples. The acoustic transfer function, the excitation waveform and PSD, area function, and midsagittal outline are updated frame by frame. For display option 2, the articulatory trajectories are displayed for the entire duration of the

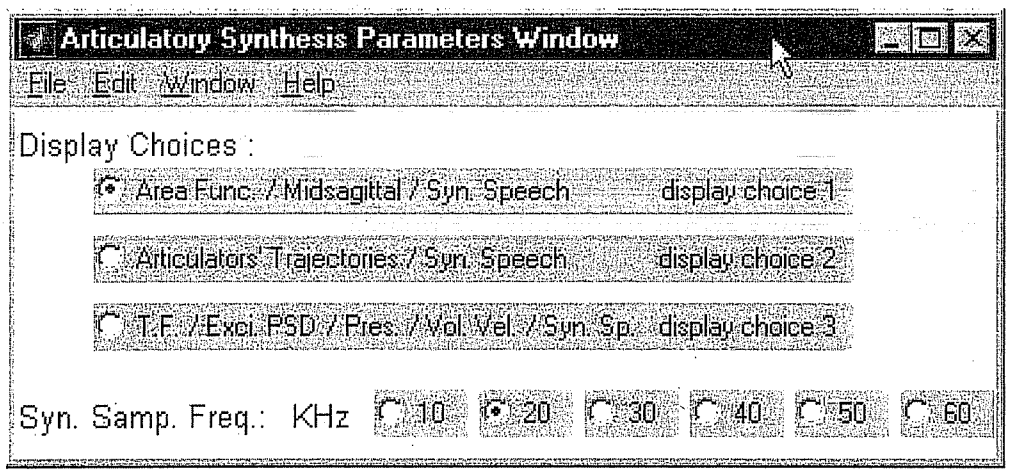


FIGURE 10.52 The articulatory synthesis parameters window.

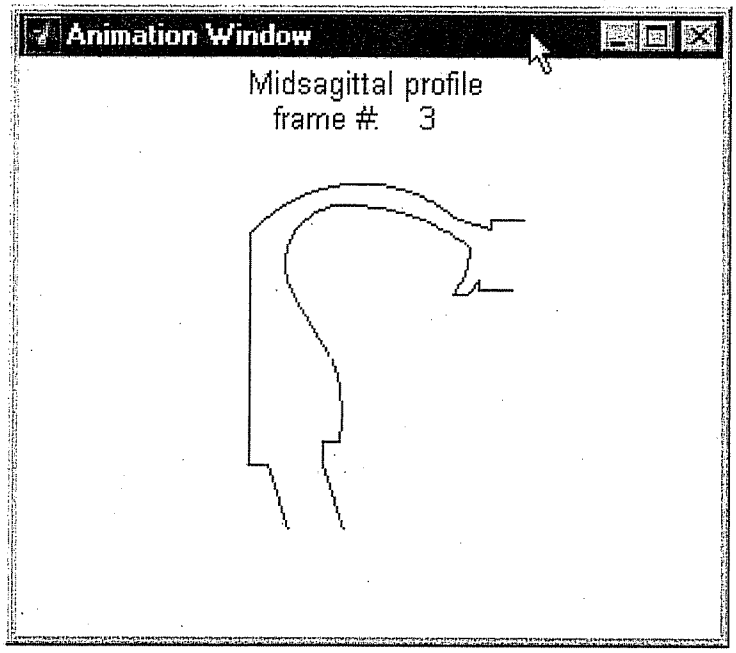


FIGURE 10.53 The animation window.

synthesis process. Display option 3 takes nearly twenty times as long to synthesize a speech sample as display options 1 or 2. Consequently, display options 1 or 2 are usually recommended. The be_total_form.for file was marked with 13 frames, then optimized, and the speech file be_total_syn_speech.dat was synthesized using display option 1. The synthesis took approximately 1.5 hours for the word be using 13 frames. Fewer frames takes much less time. Examples of the three display options appear in Figures 10.54, 10.55, and 10.56.

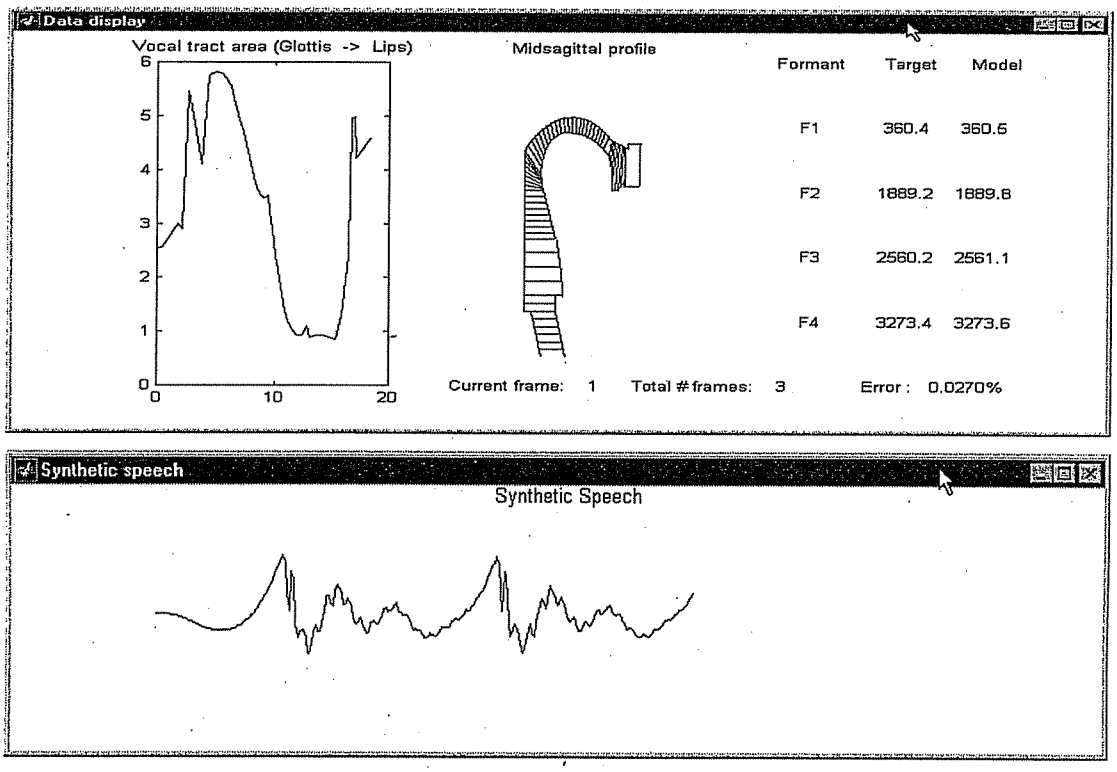


FIGURE 10.54 Display option 1.

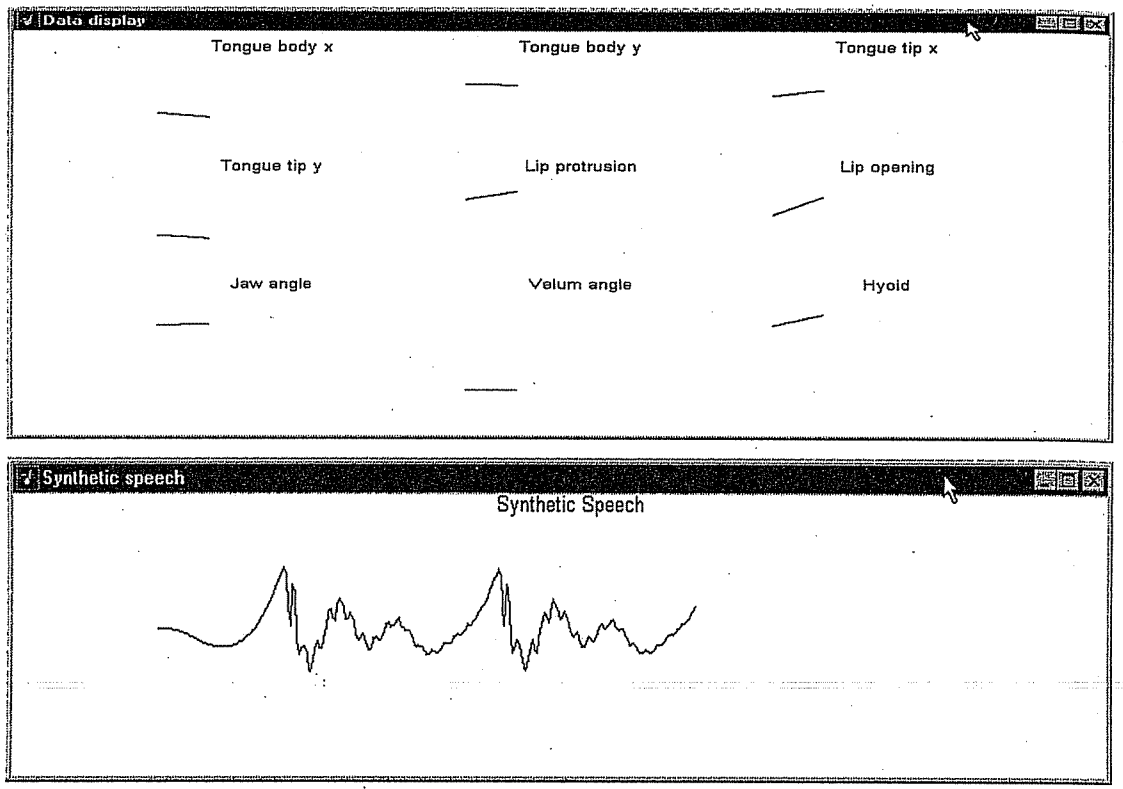


FIGURE 10.55 Display option 2.

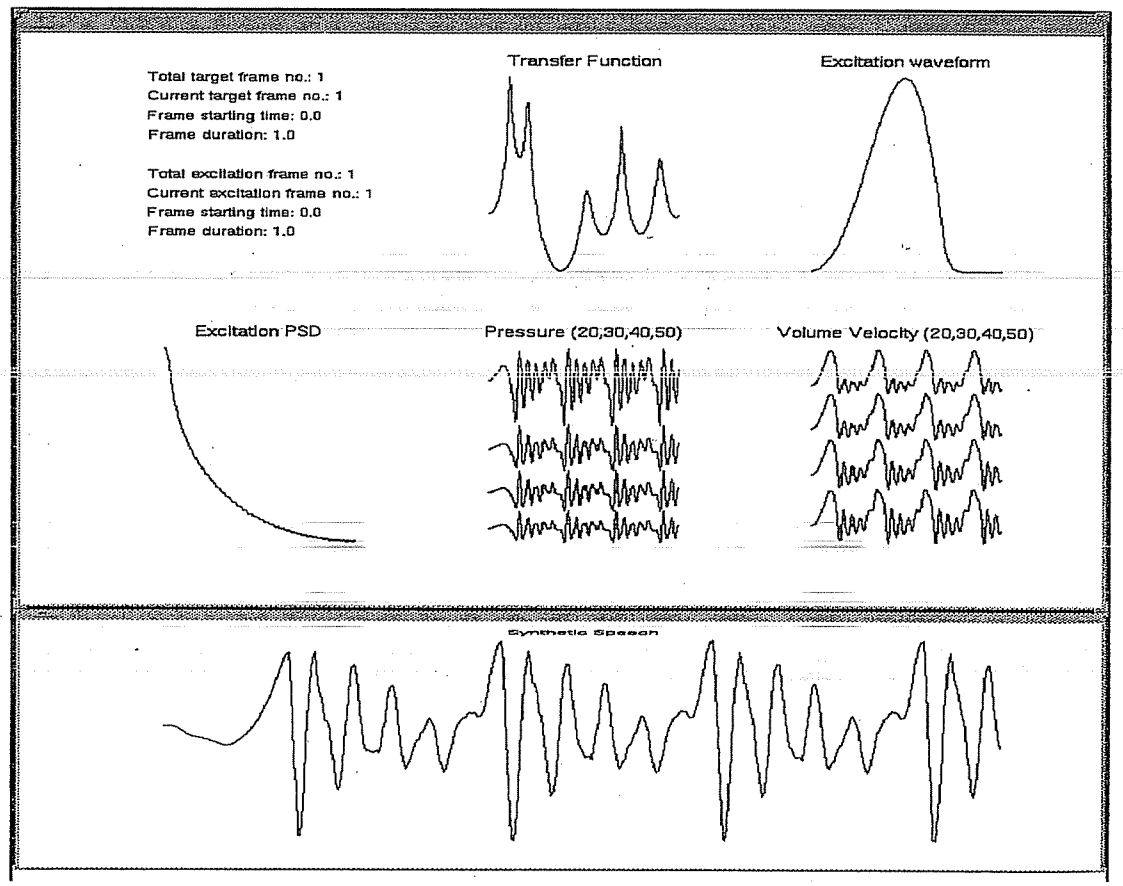


FIGURE 10.56 Display option 3.

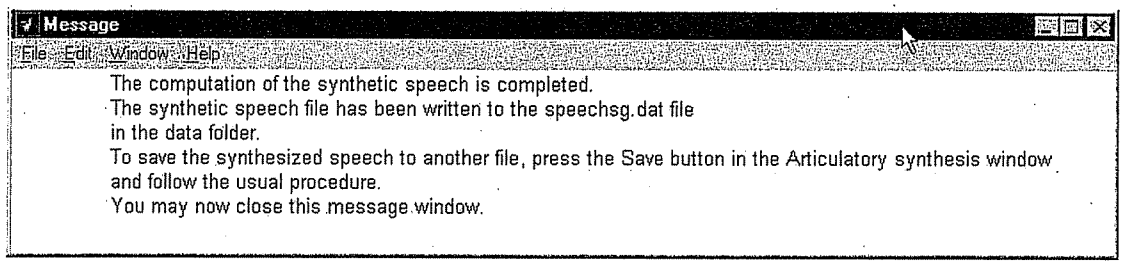


FIGURE 10.57 Message window stating that the synthesis is completed.

Upon completion of the synthesis process, a message window, shown in Figure 10.57, appears informing the user that the synthesis is completed and that the user can save the synthesized speech file. The user can print out a copy of the display (data and synthetic speech) using the Print button. The synthetic speech can be played by pressing the Play button. The speech file is saved automatically to a file called speechsg.dat. However, it is recommended that the user also save the speech file to another file by pressing the Save button, since the user may forget to rename the speechsg.dat file before performing another synthesis. A saved synthesized speech file cannot be loaded and played with this toolbox, rather the analysis toolbox or the formant_track toolbox can be used. The Clear button clears the canvas.

For each of the display options if the duration of the synthetic speech exceeds the length of the display window, then the window is cleared and the speech signal is drawn from the beginning of the synthesized speech window. For display option 3, if the duration of the synthetic speech exceeds the length of the display window, then the pressure and volume velocity waveforms overwrite the displayed waveforms.

The data display for display option 2 is not very interesting for the example shown, since the speech is a steady vowel. Consequently, the articulators do not move significantly. This display is more interesting for a sentence, such as, "Should we chase those cowboys?"

Figure 10.56 is for a vowel but is not the same data as shown in Figures 10.54 and 10.55.

10.4 SUMMARY

The operation of this toolbox entails the extraction of the first four formant tracks from a target speech file. The formant tracks are marked with frame boundaries. The shape of the articulatory vocal track is initially determined using a set of default articulatory position vectors. Inverse filtering is performed using a simulated annealing program that minimizes the error between the target formants and the model formants on a frame-by-frame basis. The inverse filtering step determines the optimized articulatory parameter vector for the target speech. Prior to synthesis, an excitation waveform is designed. The excitation can be voiced, unvoiced, or mixed excitation. The voiced and mixed excitation can include jitter and shimmer, aspiration, and a subglottal model. Mixed excitation can also include turbulence. The unvoiced excitation does not include voicing, but does include turbulence, aspiration, and the subglottal model. The synthesis phase constructs the synthesized speech waveform. The user can also view an animated movie of the midsagittal outlines for the articulatory vocal tract model. A summary of the toolbox is given in Figure 10.58.

10.5 THE DATA FOLDER

The data folder in the artm toolbox contains the following example data files, which are arranged by name: "aa" designates the vowel AA, "be" designates the word be, "m"

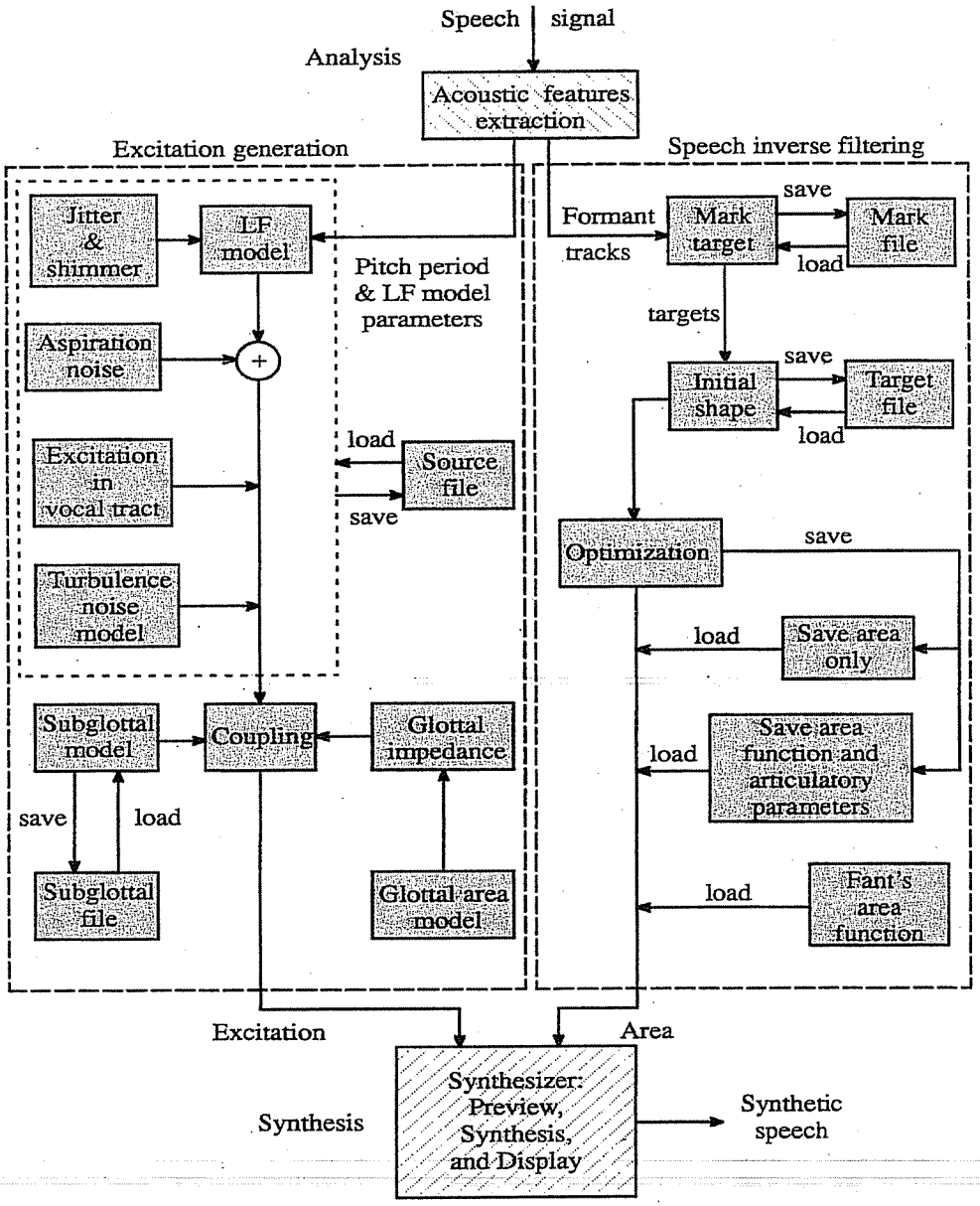


FIGURE 10.58 Outline of the articulatory speech synthesis toolbox.

designates the nasal M, "sh" designates the fricative SH, "we" designates the sentence, "We were away a year ago." The file extensions are:

*.fan	Fant area function. See Appendix 5, the Russian vowels.
*.art	Articulatory parameter file obtained by inverse filtering using simulated annealing.
*fmrk.mrk	A formant track file marked with frame boundaries.
*form.for	A formant track file obtained using the formant_track toolbox.
*.dat	A speech data file, either synthesized (syn) or natural.
aa.fan	The Fant area function file for the vowel AA.
aa_fmrk.art	The articulatory parameter file for the formants for AA.
aa_fmrk.mrk	The marked file for aa.
aa_form.for	The formant file for aa.
be_seg.art	An articulatory parameter file for a segment of the word be. Not optimized.

be_seg_fmrk.mrk	The marked file for be_seg.
be_seg_form.for	The formant file for be_seg.
be_seg_opt.art	The optimized articulatory parameter file for a segment of the word be.
be_seg_syn_speech.dat	The synthesized speech for the be_seg file using the be_seg.art file.
be_total.art	An articulatory parameter file for the total word be. Not optimized.
be_total_1_v_js_asp.src	The exicitation file used to synthesize be_seg_syn_speech.dat and be_total_syn_speech.dat. The file has one frame, voiced, jitter and shimmer, and aspiration noise.
be_total_fmrk.mrk	The marked file for be_total.
be_total_form.for	The formant file for be_total.
be_total_syn_speech.dat	The synthesized speech for be_total.
ex1_v_.src	An excitation file, one frame, voiced.
ex2_v_.src	An excitation file, two frames, voiced.
ex32_v_src	An excitation file, 32 frames, voiced.
m.fan	The Fant area function file for the nasal M.
sh.fan	The Fant area function file for fricative SH.
speechsg.dat	The speech file that is automatically saved after synthesis.
we_.art	The optimized articulatory parameter file for a segment of the sentence, "We were away a year ago."
we_fmrk.mrk	The marked file for we.
we_form.for	The formant file for we.

The following files are synthesized speech for various conditions.

avocal00.dat	AA with the excitation at the glottis (vocal tract section 00).
avocal35.dat	AA with the excitation at vocal tract section 35.
avocal50.dat	AA with the excitation at vocal tract section 50.
shcent54.dat	SH with the excitation noise source at the center of the constriction, section 54.
shdi5254.dat	SH with the excitation noise source distributed from section 52 to 54.
shdown59.dat	SH with the excitation noise source downstream from the constriction, section 59.
shup53.dat	SH with the excitation noise source upstream from the constriction, section 53.
swc123_1.dat	Sentence, "Should we chase those cowboys?"
syn3500.dat	Two successive vowels, both AA, both with excitation at section 35. The first is with the excitation designed by the software with no prefiltering. The second excitation is prefiltered to compensate for the relocation of the source to the 35th section. The prefiltered excitation produces a sound that is almost a tone because the fifth and sixth formants are introduced as artifacts by the prefiltering. Thus, the prefiltering needs to be improved.
wwwvoc00.dat	Sentence, "We were away a year ago," with excitation at the glottis (section 00).
wwwvoc35.dat	Sentence, "We were away a year ago," with excitation at section 35. The excitation has not been prefiltered to compensate for the relocation of the source.
wwwvoc50.dat	Sentence, "We were away a year ago," with excitation at section 50. The excitation has not been prefiltered to compensate for the relocation of the source.

To listen to and view the synthesized speech files use the analysis toolbox or the formant_track toolbox.

PROBLEMS

10.1 In Chapter 6 the synthesis tasks were primarily vowels. For this chapter, the tasks are to synthesize words and sentences. The purpose of this problem is to become familiar with the various options discussed in this chapter. To do so, you are asked to repeat some of the examples used in the text. Use the formant_track toolbox to obtain the formant tracks for the word be (file b.dat). Copy the target formant file to the data folder in the artm toolbox. Call this file be_test_form.for. Close the formant_track toolbox. Use the artm toolbox to load this file. Mark the formant tracks with two frame boundaries, approximately evenly spaced over the file, using the add mark option. Save the marked file as be_test_frmk.mrk. Start the shape option. Experiment with altering the sliders to reduce the error. Try to reduce the error manually to about 7% or less for each frame. Next, select and start the optimization option. Use the default settings in the articulatory optimization setup window. Next, select and start the simulated annealing window. Use the default values. Press Start in this window. Record the time. The optimization process may take nearly 45 minutes, depending on the type of machine you are using. Upon completion of the optimization, record the time required to do the optimization. You can save the optimized articulatory parameter file as be_test_.art if desired. Next, start the excitation option and load the ex1.src file. Alter the excitation parameters if desired, draw the waveform, and save the file to be_test_.src. Start the synthesis option. Select the animation option. Only two frames will be shown. Use the default values for the articulatory synthesis parameters. Press the Start button in the articulatory synthesis window. Record the time. The synthesis may take 1 to 1.5 hours. Upon completion of the synthesis, record the time required to do the synthesis. Save the synthesized speech file as be_test_syn_speech.dat. Play the file. Does it sound like the original? Print the file. Close out the artm toolbox. Open the formant_track toolbox. Load the file be_test_syn_speech.dat. Play the file and view its waveform and compare with the original.

10.2 Repeat Problem 10.1 but use display option 2 in the articulatory synthesis parameters window. Use different file names.

10.3 Repeat Problem 10.1 but use the 60 kHz sampling frequency and display option 1 in the articulatory synthesis parameters window. Use different file names. Compare the quality of the synthesized speech obtained from both problems.

10.4 The optimization process can be speeded up by altering the parameter settings in the simulated annealing window. For example, reduce the number of cycles to 10, the number of iterations to 5, and the maximum number of evaluations to 100. These changes will not give the optimum error, but will greatly reduce the computation time. Repeat Problem 10.1 with these changes and compare the two synthesized speech files. Be sure to use different file names. Note the errors and compare the location of the model formants with the target formants.

10.5 Repeat Problem 10.1 but change the excitation to use be_total_1_v_js_asp.src. Compare the two synthesized files. Be sure to use different file names.

10.6 Repeat Problem 10.1 but change the excitation to unvoiced. You will have to design an excitation file. Does the synthesized speech sound like whispered speech?

10.7 Use the analysis toolbox to obtain a segment "We were" from the sentence, "We were away a year ago." Repeat Problem 10.1 using the optimization process, and using the voiced excitation ex1.src. Compare the synthesized file with the original.

10.8 Repeat Problem 10.1 but synthesize the sentence, "We were away a year ago." This may take some time (see Appendix 11-D). Use 20 frames to mark the formant track file. Use display option 1 and then repeat using display option 2 so that you can observe the changes in the articulator movements.

10.9 Repeat Problem 10.8 but synthesize the sentence, "Should we chase those cowboys?"