CHAPTER 7

VOCOS—A VOICE CONVERSION TOOLBOX

7.1 INTRODUCTION

This chapter describes software that can be used for speech analysis, speech synthesis, and voice conversion, also known as voice transformation or voice manipulation. The software provides graphical user interfaces (GUIs) with various features to assist the user in speech analysis and synthesis tasks. For example, the user can select analysis algorithms, inspect and correct analysis derived parameters, align the acoustic parameters of two speakers with different speaking rates using dynamic time warping (DTW), and synthesize speech. Speech synthesis can mimic the voice of a speaker or convert the voice of one speaker to sound like that of another. Data displays are provided to assist the user in judging the correctness of the analysis results. The primary acoustic features that are measured include voiced/unvoiced (voice type) classification, pitch and gain contours, formant frequencies and bandwidths, and the shape and type of glottal excitation waveform. An overview of the major characteristics of the software is shown in Figure 7.1. The software system is called VOCOS for Voice Conversion System. The software is to be installed in a subdirectory in MATLAB in a manner similar to that described for the analysis system. Additional details are provided in Chapter 6, Appendices 7 and 8, and Hsiao (1996).

7.2 MAIN FUNCTION

The analysis algorithms are based on a fixed-frame linear prediction (LP) analysis method. To start the software change directory to the VOCOS directory and type main in the MATLAB command window. The Main Function window shown in Figure 7.2 appears. The Analysis button provides algorithms for the analysis of speech data. The Correction button lets the user correct an analyzed speech data file. The Modification button allows the user to modify parameter values of an analyzed speech file. The Synthesis button provides a method for synthesizing speech using the parameters obtained during speech analysis. There are two types of synthesis: linear prediction (LP) and formant. The selection of the type of synthesizer is automatic and determined by the type of vocal tract model selected by the user during the analysis phase. The Close button closes out the Main Function window.

7.3 SPEECH ANALYSIS—LINEAR PREDICTION (LP)

Upon pressing the Analysis button, the window shown in Figure 7.3 appears. The topmost button, Specification, is used to specify the desired analysis parameters, which are available in the Specification window, shown in Figure 7.4. This window contains a set of default

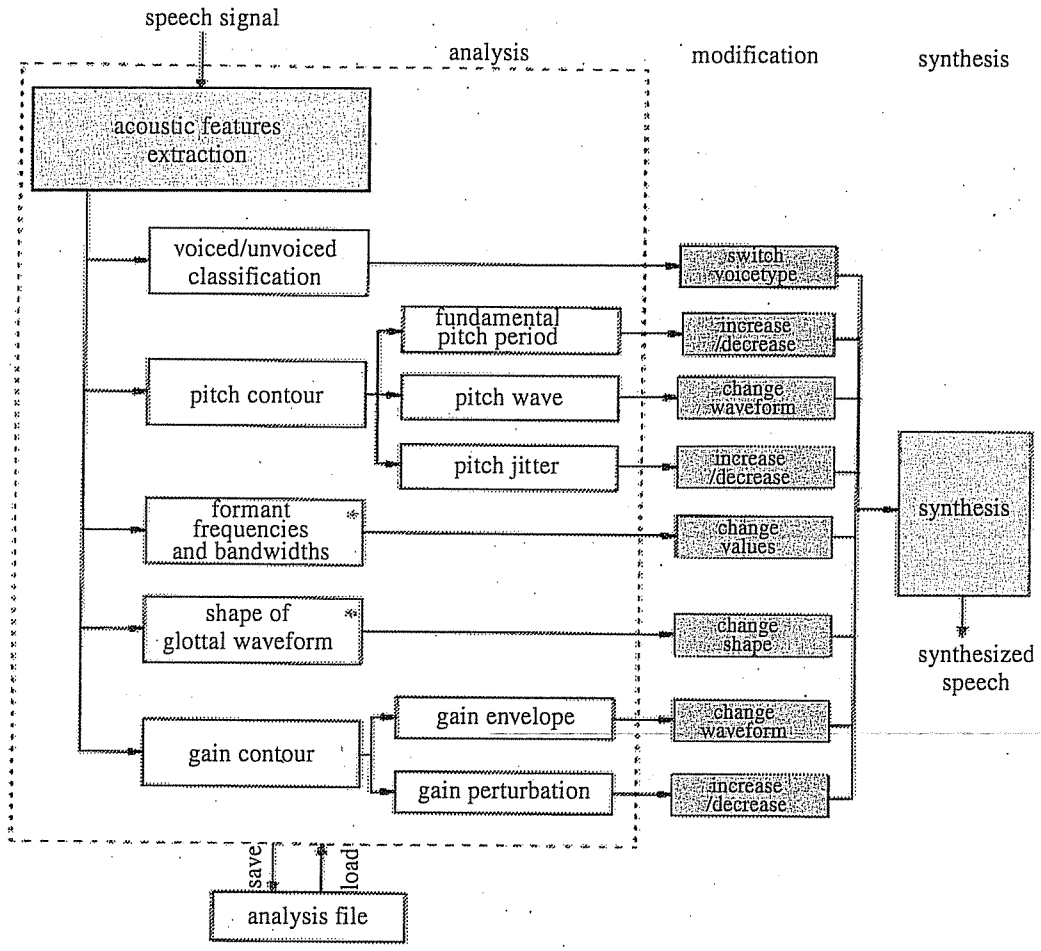

FIGURE 7.1 Voice conversion software system. The blocks with an asterisk are not available for all analysis parameter settings.

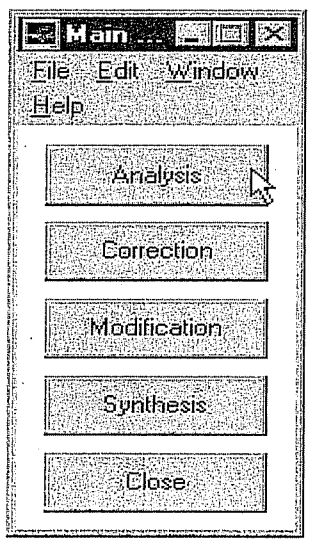

FIGURE 7.2 Main function window.

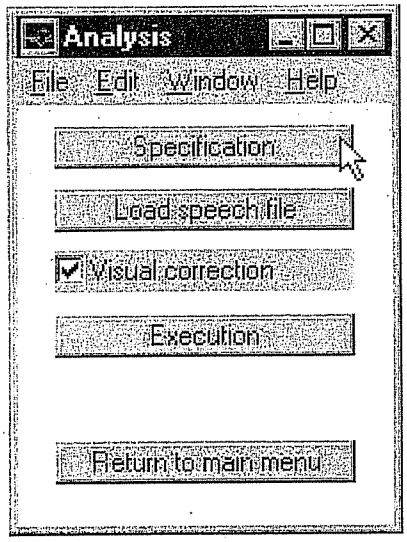

FIGURE 7.3 Analysis window.

values that can be reset using the Default button. There are two source models: polynomial and simplified LF. The number of formants can be selected as well as the window frame length and the window overlap. The two vocal tract models are linear prediction and formant. The order of the LP model can be selected. After the desired parameters have been set, press the Apply button, otherwise reset the values to the default by pressing the Default button. Finally, press the Return button to close out the Specification window and return to the Analysis window for the next selection.

Next, load a data file for analysis by pressing the Load Speech File button. This brings up the window shown in Figure 7.5, which shows the contents of the `spedata` subdirectory. This directory is designed to contain speech data files for analysis or voice conversion. If the desired data file is not in this directory, then change directory to the desired location. In Figure 7.5 the `m11.dat` file is selected and loaded. This file is displayed as the Input Signal in Figure 7.6. This is the sentence, "We were away a year ago," spoken by a male speaker. This sentence is to be analyzed. Remember, do not close the Input Signal window or the Main Function window until the analysis is complete, since all data will be lost. To proceed, the user presses the Execution button in Figure 7.3. Note that the visual correction option for analysis is checked as the default. This is almost always the desired option. On occasion, the user can uncheck this option. For example, when an analysis of familiar data is being performed and the user knows that visual correction is not necessary or is to be omitted.

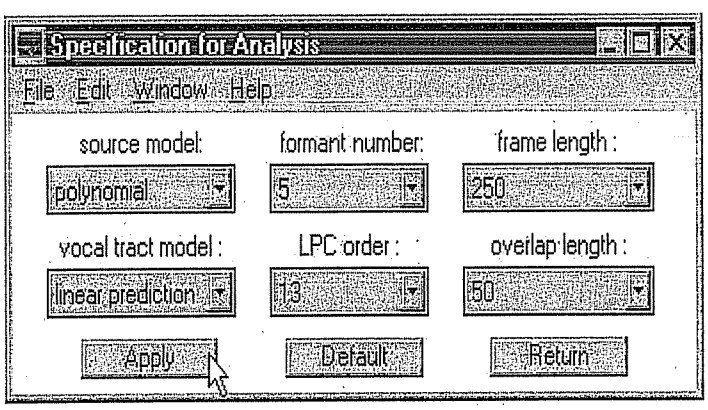

FIGURE 7.4 Specification window for analysis.

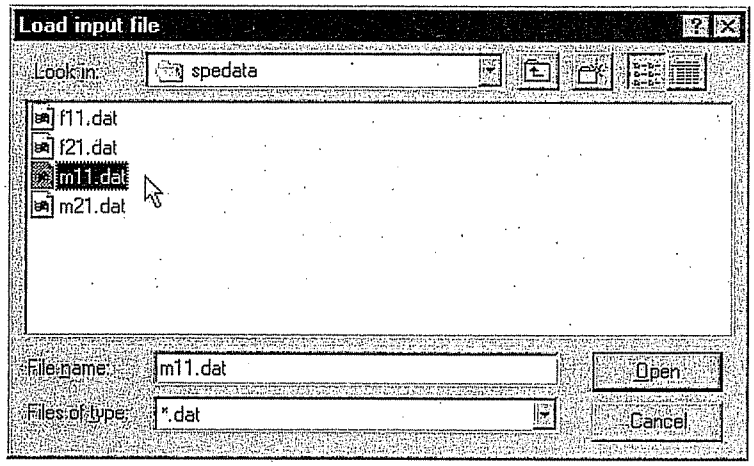

FIGURE 7.5 Load input file window.

The data files stored in the spedata directory are dat speech files. The data are sampled at 10 kHz. The prefix f/m refers to female/male speaker. The next digit is the speaker identification number, and the following digit is the sentence identification number. Thus, m11.dat is the speech file for male speaker 1, sentence 1, which is, "We were away a year ago." The file names have been shortened for this chapter only as a convenience, the reasons for which will become apparent later.

After starting the execution of the software, a message window appears stating that the program is working. When the first stage of calculations is completed, a second message window appears stating that the voice types (voiced/unvoiced) are classified and that the user can inspect the results. In addition, the action window shown in Figure 7.7 appears, allowing the user to view the results of the first stage of calculations or to continue. Upon selecting the Voice Type Inspection option, the results shown in Figures 7.8 and 7.9 appear. These results are obtained via an automatic voice type classification algorithm that can make mistakes, thereby, introducing unwanted errors. Figure 7.8 is an action window that allows the user to scroll through the data to view the results of the voice type classification algorithm. Press the > (<) button to scroll the data to the right (left). This figure shows the voice type classification at the top of the data as either V or U for voiced or unvoiced, respectively. The

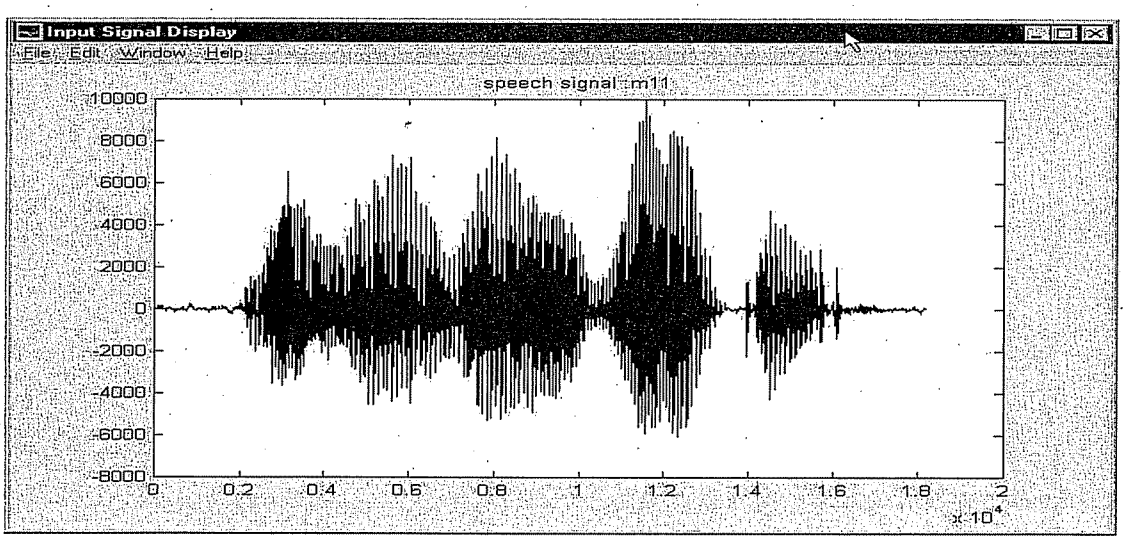

FIGURE 7.6 Input signal for analysis.

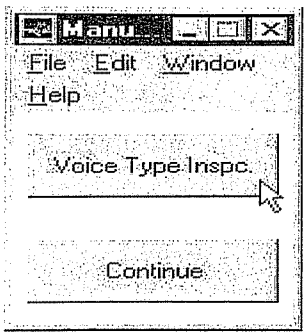

FIGURE 7.7 Voice type inspection window.

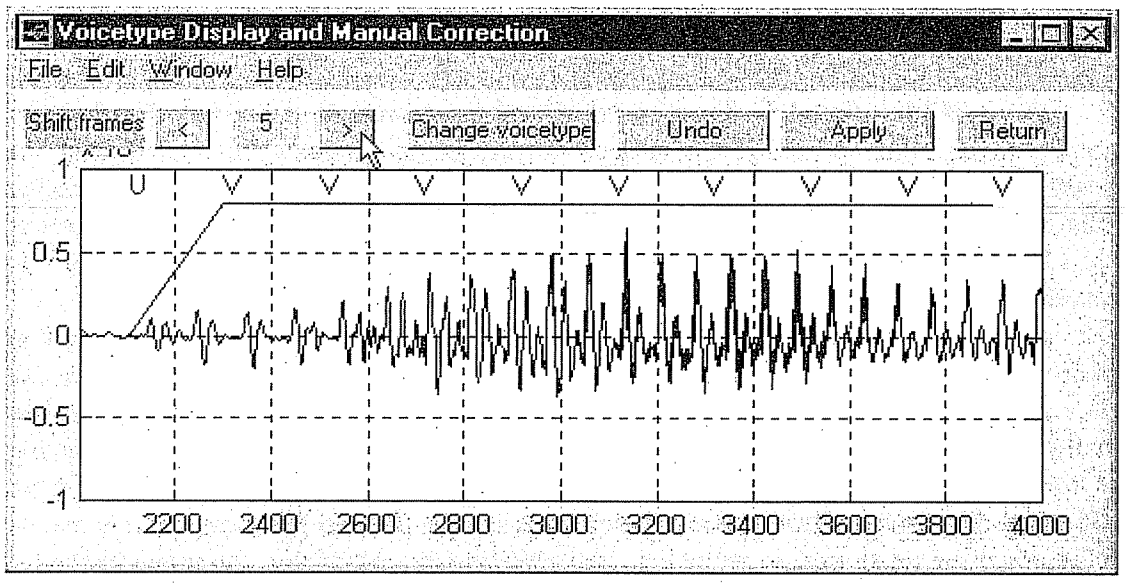

FIGURE 7.8 Voice type display and manual correction window.

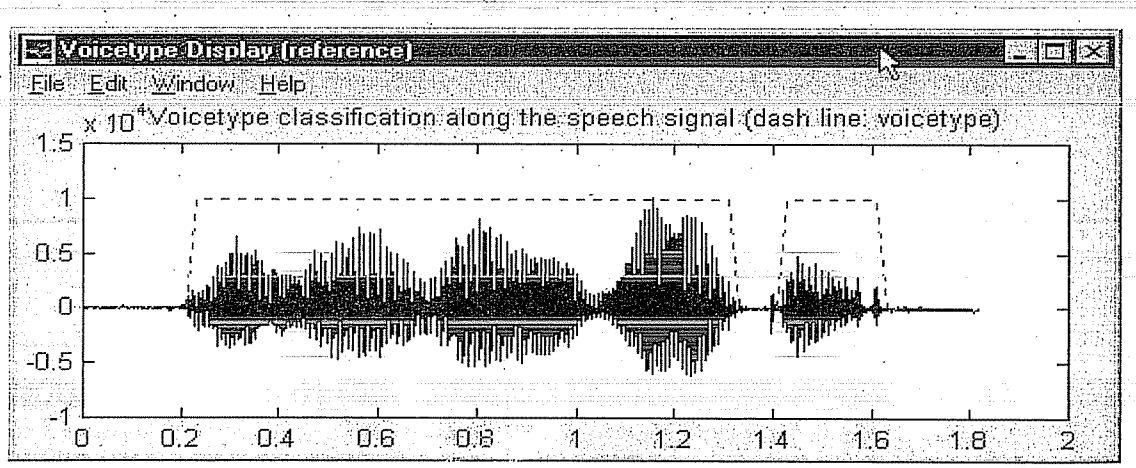

FIGURE 7.9 Voice type display (reference) window.

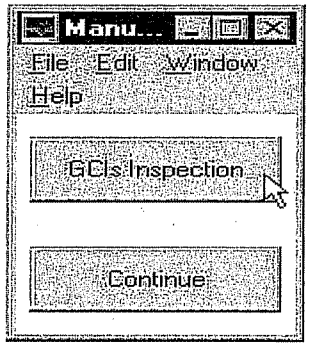

FIGURE 7.10 GCI inspection window.

inclined, solid line indicates the onset of voicing and remains constant as long as voicing is present. An overview of the V/U classification for the complete data file is shown in Figure 7.9, where the dashed line superimposed over the data indicates the voiced data regions. As the user scrolls through the data in Figure 7.8, errors can be found and corrected as follows. If an error is found, say a frame is marked by the algorithm as voiced (V) and it should be unvoiced (U), then press the Change Voice Type button. The mouse cursor changes to a cross hair. Move the cross hair to the V and press the left mouse button. The V changes to a U and the solid line changes to zero, indicating unvoiced. A V > U indication is printed under the V to remind the user that a change has been made. The cross hair changes back to the usual cursor. To undo this operation, press the Undo button. To retain the change, press the Apply button. Perhaps the best procedure to follow is to make all the desired changes before the Apply button is pressed. However, changes can be made after the Apply button is pressed by following the above steps. Once the Return button is pressed and the analysis continues, changes cannot be made until the analysis is completed. This will be discussed later.

The analysis is continued by pressing the Continue button, whereupon a message window appears informing the user that the glottal closure instants (GCIs) are located and that the results can be inspected, in a manner as described previously. Upon selecting the GCIs Inspection button in Figure 7.10, Figures 7.11 and 7.12 are displayed.

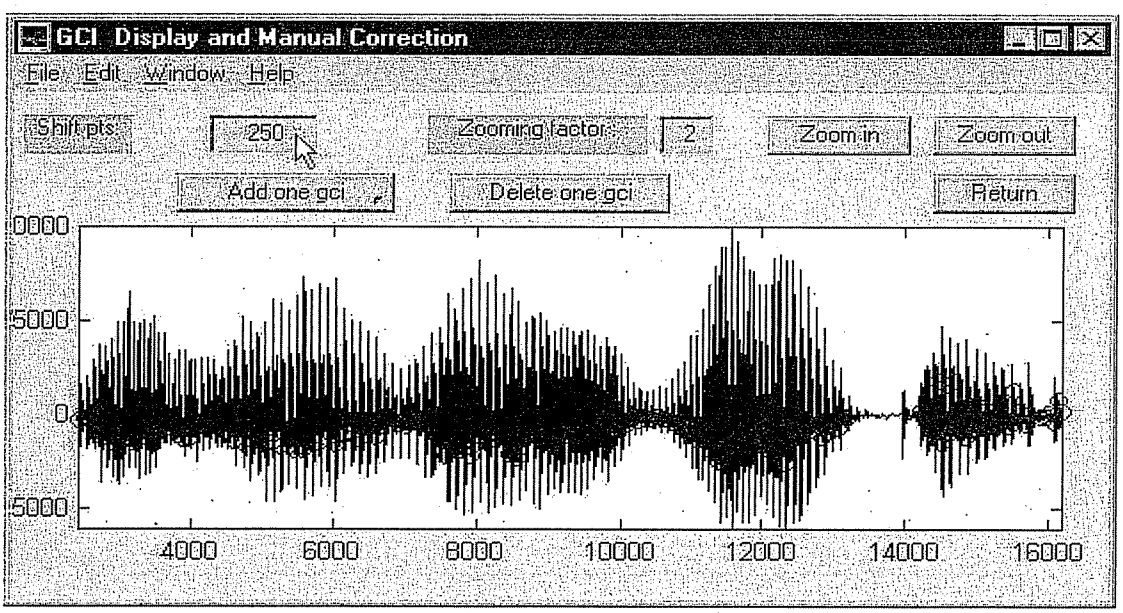

FIGURE 7.11 GCI display and manual correction window.

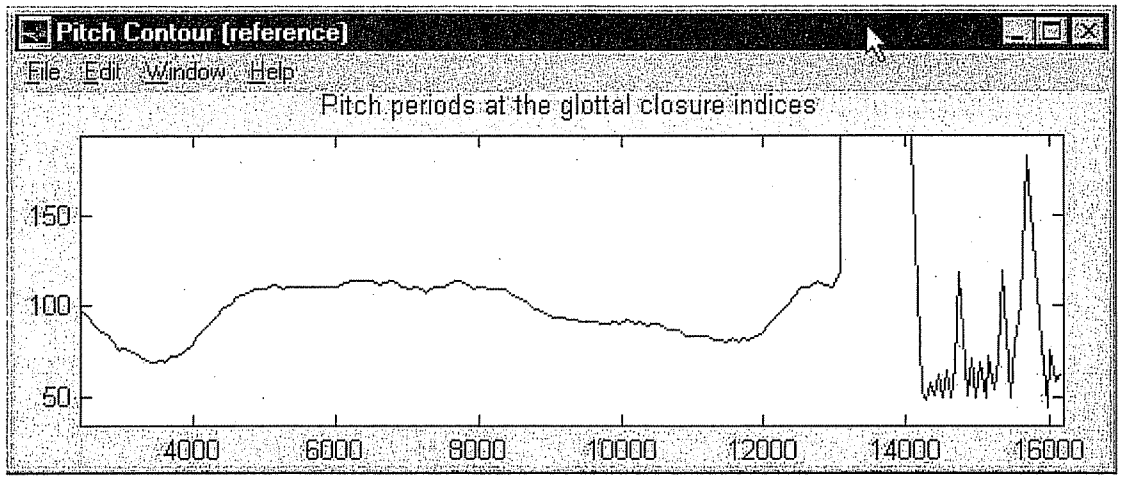

FIGURE 7.12 Pitch contour (reference).

The action window in Figure 7.11 allows the user to correct the algorithm detection errors of the glottal closure instants (GCIs). At the upper left is a panel where the user can set the number of points that the data are shifted each time. The zooming factor default setting is 2, but this can also be changed. The user can zoom in or out. The buttons on the next row are used to add or delete a GCI one at a time. Figure 7.12 shows the pitch contour calculated using the GCIs. This contour appears reasonable, except beyond point 14,000, which is the beginning of the word "go." We can see sharp changes in the pitch contour in the region 14,000 to 16,000, which should not be present, indicating some errors in the detection of the GCIs in this region. To correct these errors, we zoom in on Figure 7.11, as shown in Figure 7.13. To accomplish this, proceed as follows. In Figure 7.11, move the mouse cursor to the zoom factor panel and press the Zoom in Button. The mouse cursor changes to a cross hair. Move the cross hair to a location near 14,000 on the data and press the left mouse button. The data zooms in a factor of 2 and is replotted. Repeat this step one or more times. Note that once the zoom-in action is activated, that markers < and > appear on either side

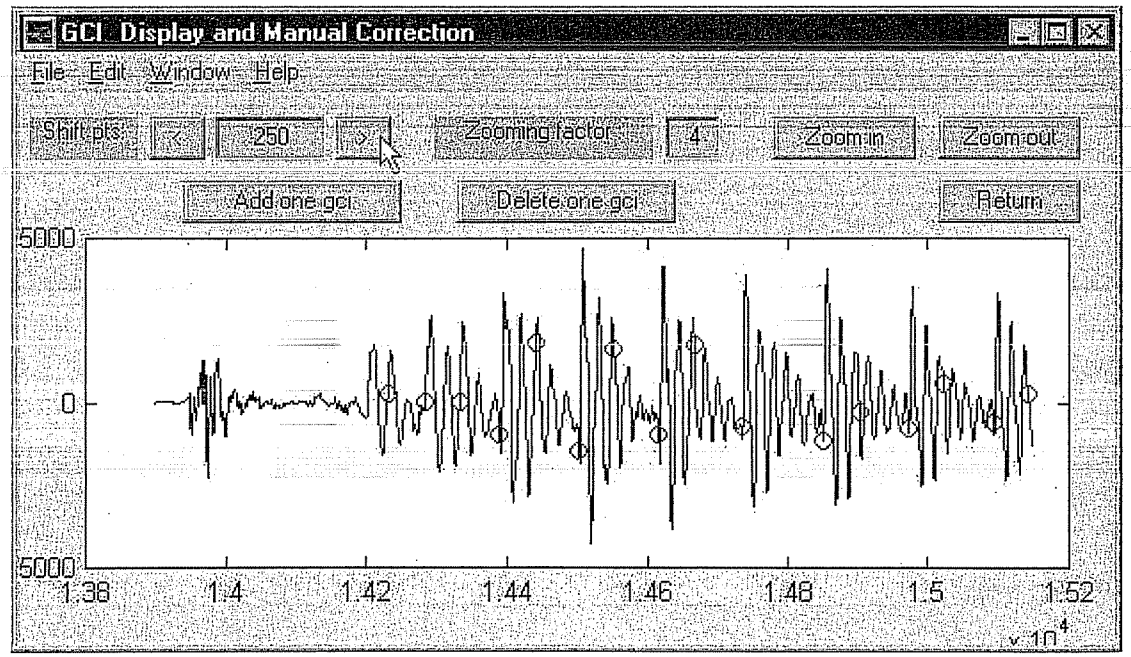

FIGURE 7.13 GCI display and manual correction window zoomed in and scrolled to 14,000.

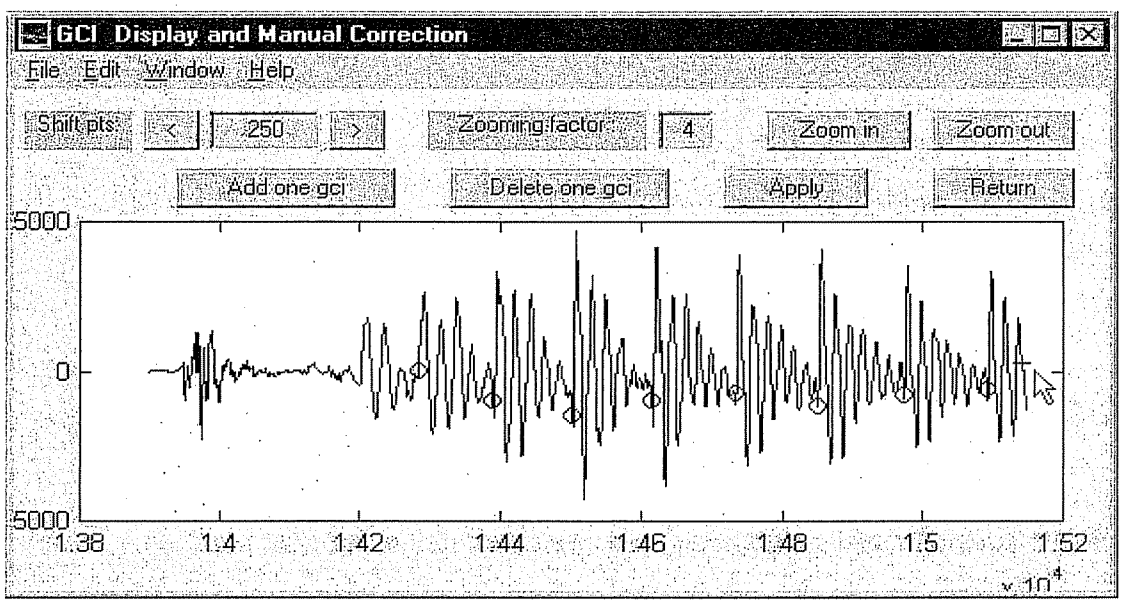

FIGURE 7.14 GCI display after correcting the GCI errors in Figure 7.13.

of the shift points panel. Scroll to the right until you reach the point 1.4×10^4 as shown in Figure 7.13. Observe that there are errors in the GCIs detection at almost every other open circle on the data. We correct these errors by removing the unwanted GCIs by pressing the Delete One GCI button, whereupon the mouse cursor becomes a cross hair. Move the cross hair to the circle to be deleted and press the left mouse button. Repeat this process until all errors are corrected. An example of the correction process is shown in Figure 7.14, where the unwanted GCIs shown in Figure 7.13 are deleted. If a GCI is to be added, follow a similar procedure. Each time a GCI is added or deleted, the Pitch contour is updated. Scroll through the data as needed. Upon making the necessary corrections, the final Pitch contour appears as shown in Figure 7.15. Note that the contour is now smoother in the vicinity of 14,000 to 16,000. To save the corrections, press the Apply button in Figure 7.14, followed by a press of the Return button. Then press the Continue button in Figure 7.10.

A message window appears stating that calculations are being performed. Another message window appears after the calculations are complete, and states that the results can be saved. The Analysis window changes to that shown in Figure 7.16, where a Save Analysis

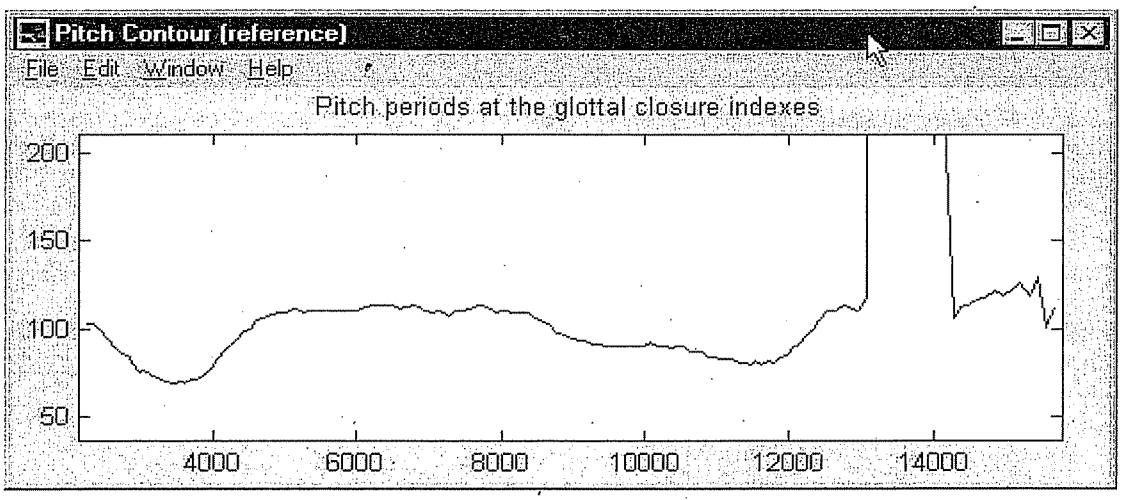

FIGURE 7.15 Pitch contour after GCI corrections.

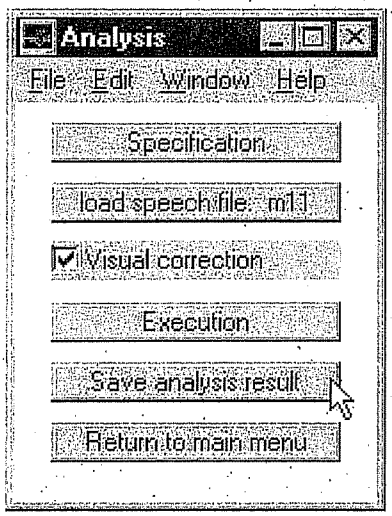

FIGURE 7.16 Analysis window upon completion of data analysis.

Result button appears. The user can save the results to a file. However, if the save option is not selected, the results of the analysis are retained in memory and the user can now do additional corrections, modifications, or synthesis. Pressing the Save File button brings up the Save window shown in Figure 7.17. The file name appears as m11.mat because we analyzed a data file named m11.dat. The file name is shown changed to m11_poly_lp_anal.mat to illustrate that the user can change the name as desired. The extension to this file is mat rather than dat, because a mat file is designed for storing a number of parameter vectors. Note, however, that the prefix, m11, is not important because the analysis files store the speech data as a vector along with the parameter vectors calculated via the analysis software. Press the Apply button in Figure 7.17. The results of the analysis are saved to the desired file. A message window appears stating that the results are saved and that the user can return to the Main window. If the user presses the Return button, both the Analysis menu window and the Input Signal window close. The user can analyze a new data file, make corrections or modifications to the data just analyzed, or synthesize speech using the parameters in the file m11_poly_lp_anal.mat. Suppose we synthesize a speech file.

7.4 SYNTHESIZE

Speech synthesis attempts to mimic the voice of the original speaker using the parameters measured in the analysis phase. Press the Synthesis button in the Main window. The

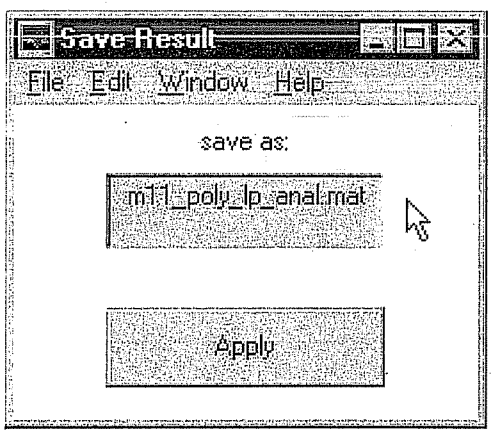

FIGURE 7.17 Save window.

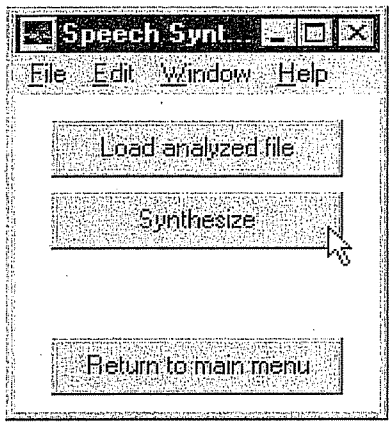

FIGURE 7.18 Synthesis window.

Synthesis window shown in Figure 7.18 appears. The available options are to Load a previously analyzed data file, such as m11_poly_lp_anal.mat, or Synthesize, or Return. In this case, the user can press the Synthesize button, since the analyzed data are still in memory. On doing so, a message window appears stating that the synthesis process is underway. After the calculations for synthesis are completed a message window appears stating that the synthesis is completed and that the synthesis file can be saved. In addition, the synthesized and original data waveforms are plotted in Figure 7.19. Recall that the original data file is m11.dat. Figure 7.19 is an action window that lets the user play the data files, plot the spectrograms of the data as in Figure 7.20, or replot the waveforms. The Return button closes the figure and returns to the Synthesis window. The synthesized data can be saved to a file by pressing the Save Synthesized Speech button in Figure 7.21, which appears after the synthesis process is completed. This save option is similar to that described previously, except that the file saved is a dat file. For example, in this case the saved data file name is m11_poly_lp_syn.dat.

Note that the software stores the analyzed parameters, e.g., source model (polynomial or LF), vocal tract model (LP or formant), and so forth, as well as the speech data loaded as the original data file. The type of speech synthesis is selected automatically as either linear prediction (LP) or formant depending on whether the analysis phase used the LP or formant vocal tract model, respectively.

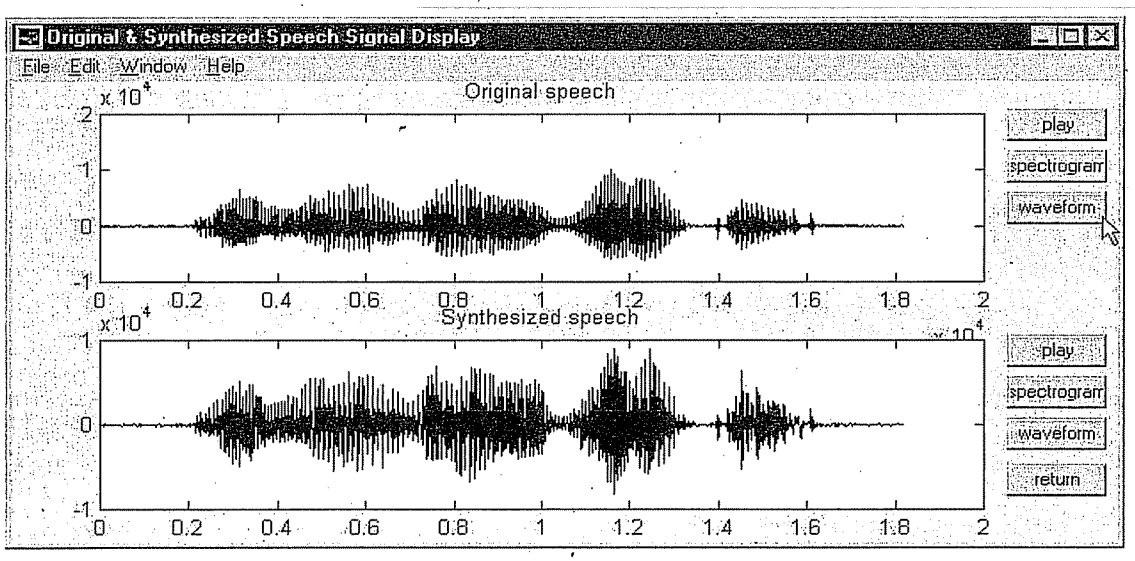

FIGURE 7.19 Original and synthesized speech signal display window showing waveforms.

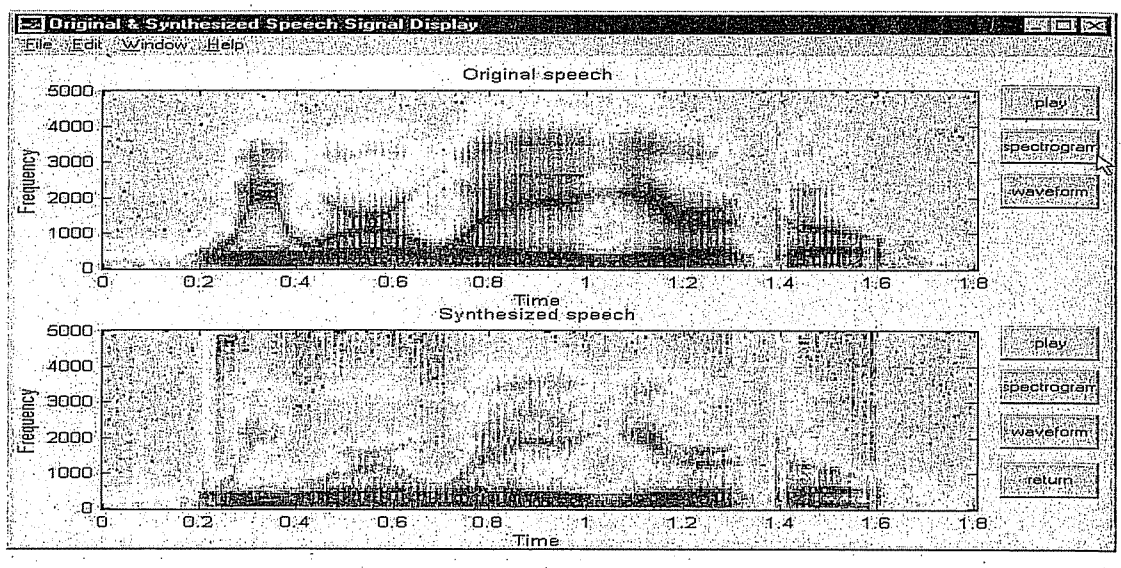

FIGURE 7.20 Original and synthesized speech signal display window showing spectrograms.

7.5 SPEECH ANALYSIS—FORMANT

The previous analysis procedure used the polynomial model as the source and the vocal tract linear prediction model with the other parameters set to the default values. In this case, the formant frequencies are not calculated. To illustrate this aspect of the software, return to the Main menu window, select Analysis, and set the source model to simplified LF and the vocal tract model to formant. Leave the other parameters set to their default values. The source model does not affect the vocal tract model calculations, so it could have remained set at polynomial. Load the m11.dat file again for analysis. Press the Execution button. The voice type and GCI analyses are performed as described previously. The third and last analysis in this case is the formant analysis, which can be inspected in Figures 7.22 and 7.23. These two figures are similar to those for the voice type and GCI analysis results. Figure 7.22 is an action figure that allows the user to alter the circled formant frequency values as follows. First, Zoom In on the data by pressing the Zoom In button. The mouse cursor changes to a cross hair, which is to be moved to the desired location for a correction. Click the left mouse button, and the data are zoomed in. While corrections can be made to the formant values without zooming in, it is easier to observe the results in the zoom-in mode. The user can scroll the data by pressing the < or > buttons. Next, press the Select Formant to Correct

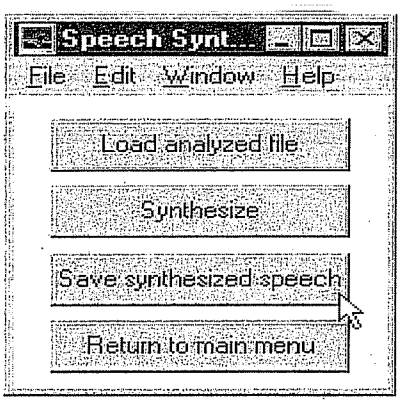

FIGURE 7.21 Synthesis window showing save option.

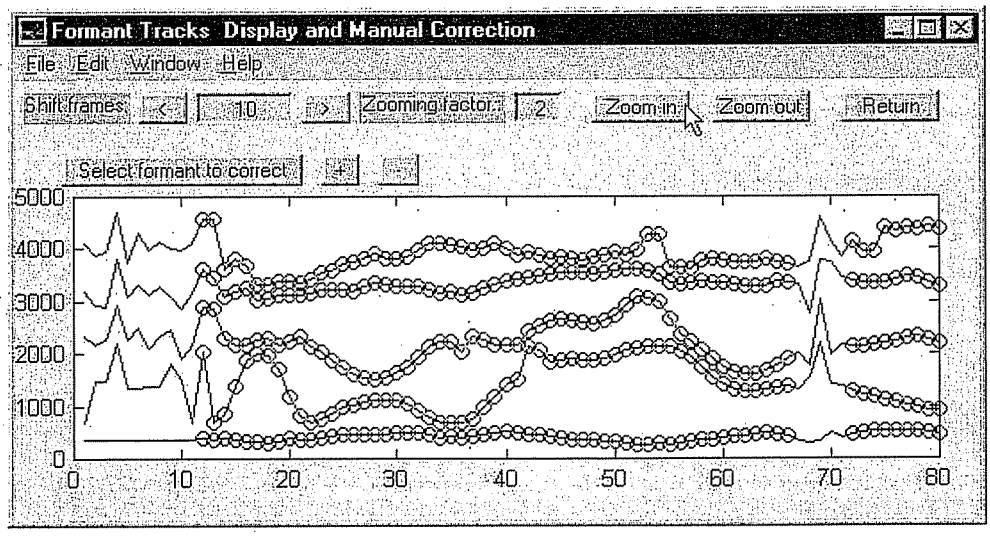

FIGURE 7.22 Formant tracks display and manual correction.

button. The mouse cursor changes to a cross hair. Move the cross hair to the circle to be corrected and press the left mouse button. An × appears within the circle and the cross hair reverts to the normal mouse cursor. If the circled formant frequency is to be reduced in value, then place the mouse cursor on the minus sign and press the left mouse button repeatedly until the circle moves down to the desired location. An example is shown in Figure 7.24. If the circled value is to be increased, then press the + button. Continue with any additional corrections as needed. To save the corrections, press the Apply button, followed by a press of the Return button. This completes the analysis and the results can be saved as described previously. The results for this analysis are saved to the file m11_lf_for_anal.mat.

7.6 CORRECTION

Pressing the Correction button in the Main menu window brings up the window shown in Figure 7.25. Note that there is no load option. The corrections are to be made to the file in memory. Assume this file is m11_poly_lp_anal.mat. Pressing any of the buttons activates the respective option. Proceed as described previously. Note that no formant information is available for this file, since we selected the linear prediction model for the vocal tract

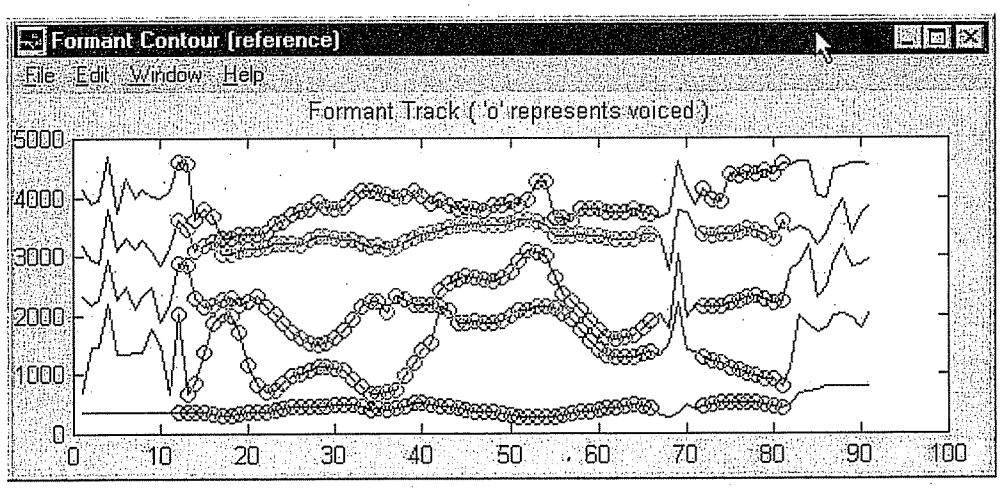

FIGURE 7.23 Formant contour reference.

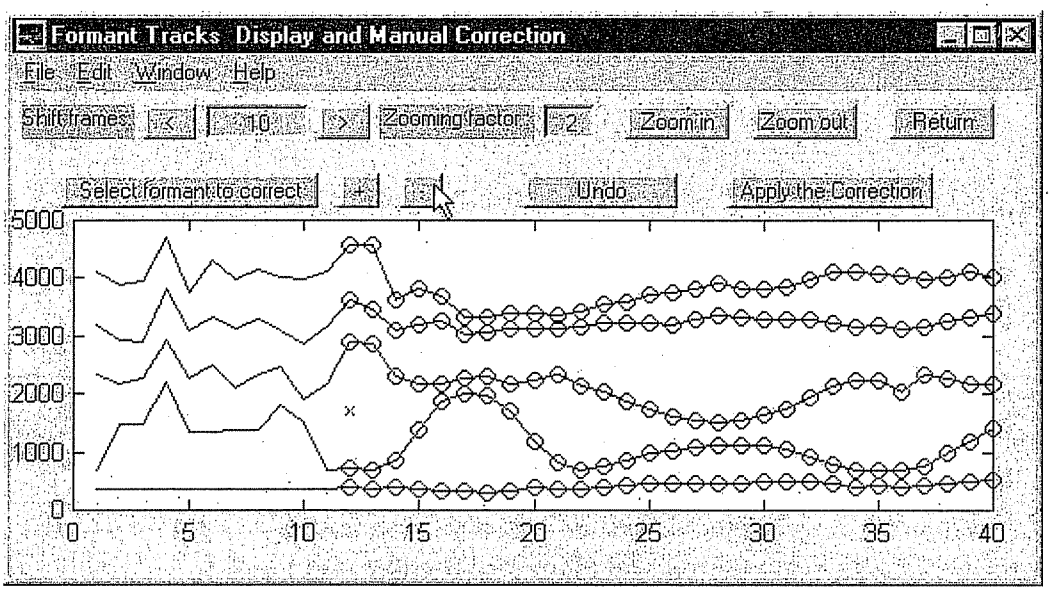

FIGURE 7.24 An example of correcting the first value of the second formant contour.

model. If we had selected the formant model, then formant information would be available. We discuss this option later. After making any necessary corrections, save the corrected file before returning to the Main window.

7.7 MODIFICATION

The Modification option window appears in Figure 7.26. The available options are described next. The Load button lets the user load a previously analyzed mat file, such as m11_poly_lp_anal.mat. However, if the file is already in memory, then the load option is not necessary.

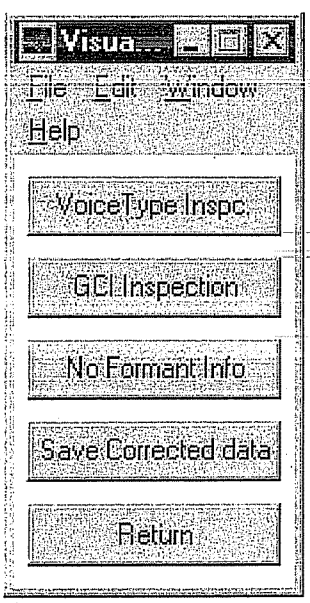

FIGURE 7.25 Correction window.

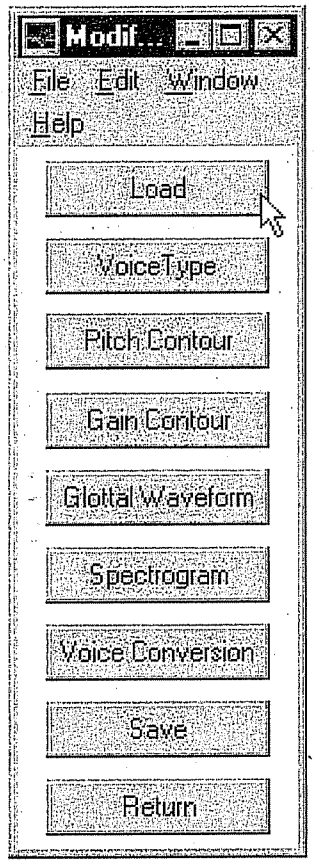

FIGURE 7.26 Modification menu window.

7.8 VOICE TYPE

Pressing the Voice Type button opens the window shown in Figure 7.27, which shows the voiced and unvoiced segments of the analyzed data as a solid line. Modifications can be made by first pressing the > marker at the upper left once, thereby moving the frame location to 2. The display automatically changes with + signs appearing on the figure, and the +

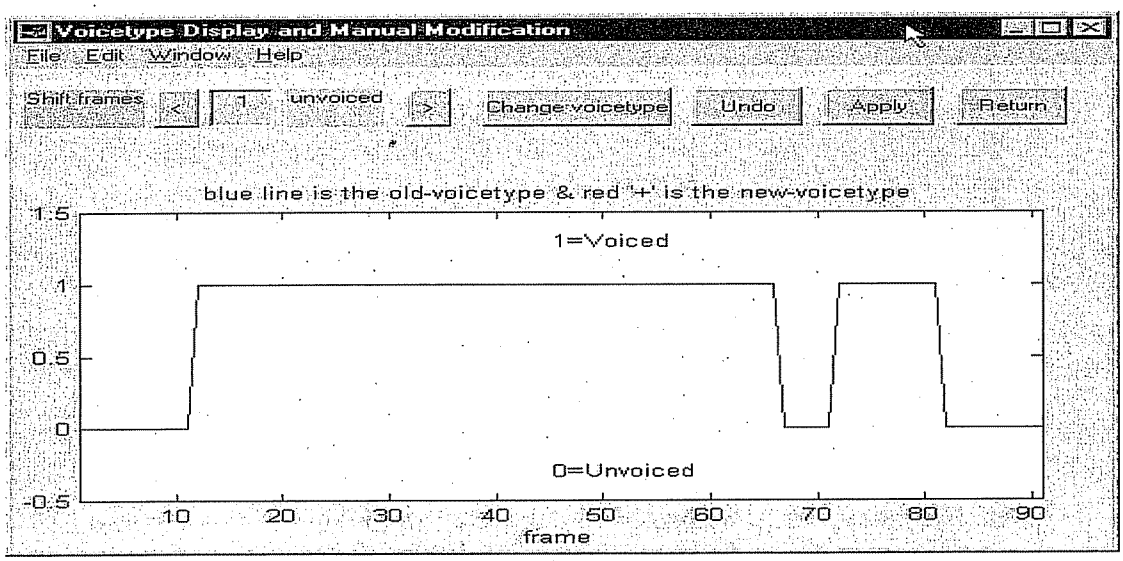

FIGURE 7.27 Voice type modification window.

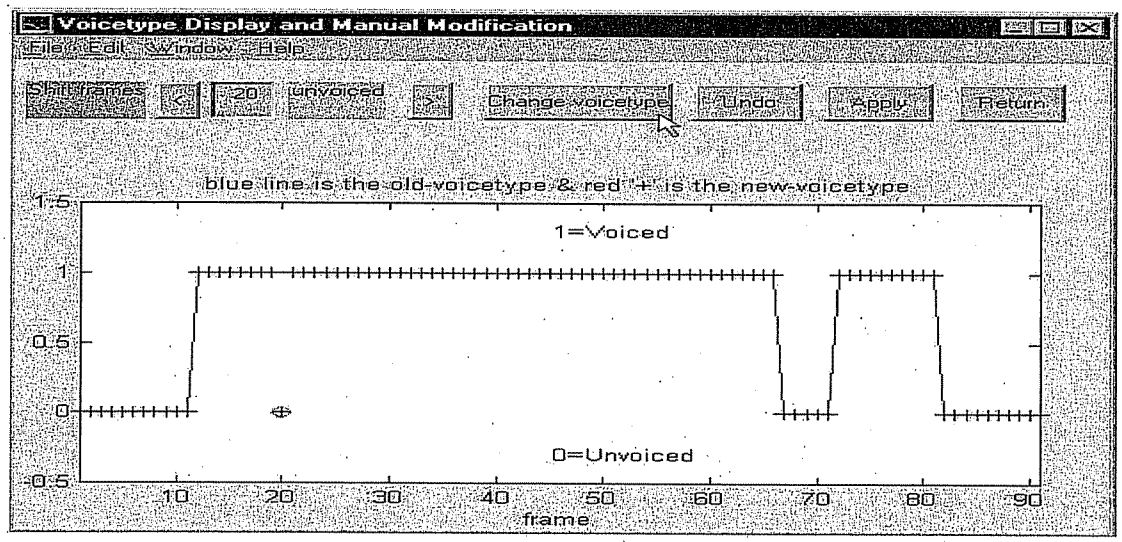

FIGURE 7.28 Voice type modification window. Frame shifted to 20 and changed from voiced to unvoiced.

sign at 2 is circled. To illustrate this, suppose we shift the location to 20 by either pressing > repeatedly or by typing in the number 20 in the panel and pressing >. Next, press the Change Voice Type button. The circled + moves from voiced to unvoiced, as shown in Figure 7.28. Pressing the Undo button can reverse this action. After all changes are made, press the Apply button followed by a press of the Return button.

7.9 PITCH CONTOUR

After pressing the Pitch contour button, three windows appear allowing the modification of the pitch contour. The left window (Figure 7.29) supplies the action buttons and sliders for altering the several factors. The upper right window (Figure 7.30) shows the pitch wave and the pitch jitter. The lower right window (Figure 7.31) displays the fundamental pitch period and the pitch contour. The pitch contour consists of three factors that can be altered separately. The fundamental pitch period is controlled by the topmost slider. Click the left side button of the slider to decrease the value, or the right side button to increase the value. The blue waveform moves down or up, respectively The jitter waveform is controlled by the lower slider and can be changed in a similar manner. It may appear that the pitch wave is altered when altering the jitter. However, this is due to a vertical scale change on the figure. The pitch wave is modified as described next.

In the middle of the left window, there are two menus and one button for modeling the pitch wave. The add knob | delete knob popup menu is used to mark or delete a critical point on the pitch wave with a knob that controls the shape of the modeled waveform. The placement of a knob on the waveform segments the waveform. The line_fit | parabola-fit(1) | parabola-fit(2) | cubic-fit | default popup menu provides five models for each segment of the pitch wave. The line_fit model draws a straight line between the two successive knobs. See MATLAB Help for the parabola-fit and cubic-fit models (help parafit and help cubfit). Briefly, the parabola-fit models determine a second-order polynomial for each segment of the waveform. The parabola-fit(1) model is an upward polynomial and the parabola-fit(2) model is a downward polynomial. The cubic-fit model determines a cubic polynomial for the pitch wave. The default model is the data waveform with no model fit to the data. The modeled wave is drawn and the knobs are marked by + signs. The draw wave button enables

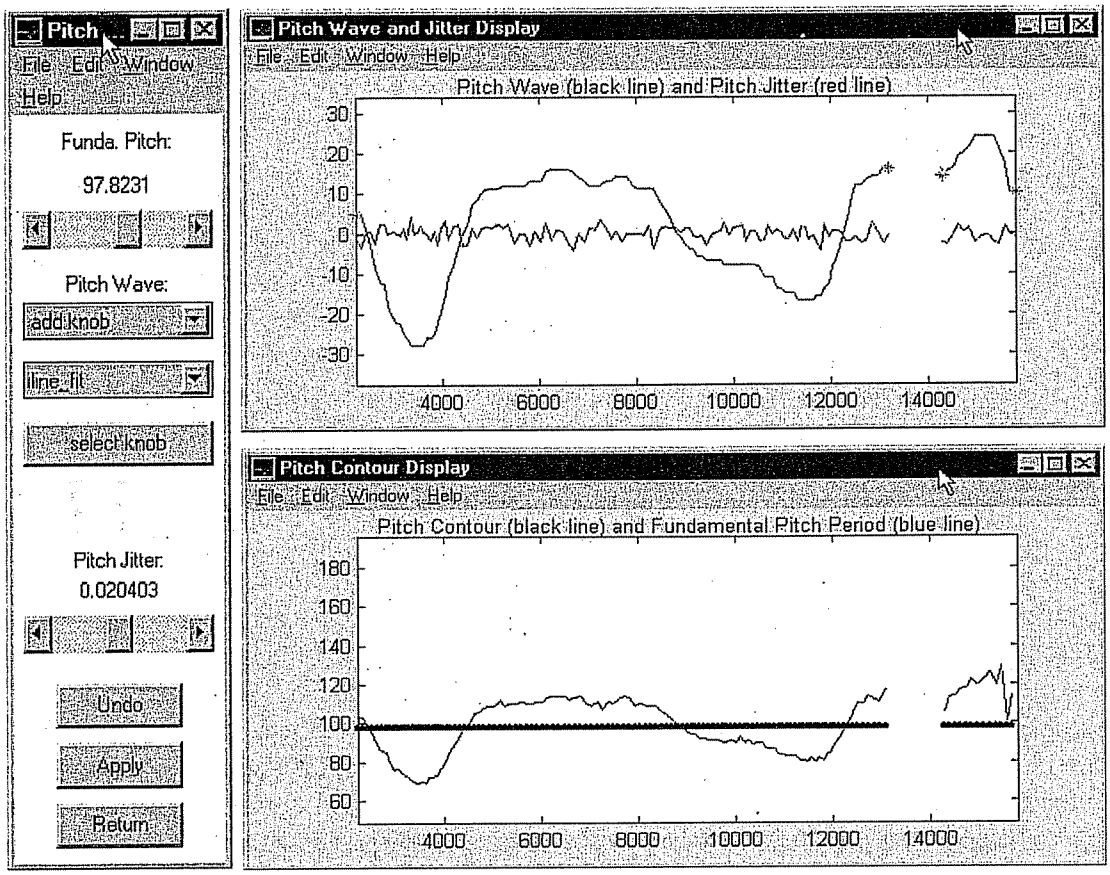

FIGURE 7.29–7.31 Pitch modification window (left). Pitch wave and jitter window (upper right). Pitch contour and fundamental pitch period window (lower right).

the knob to be moved in the vertical direction to a desired location that does not have to be on the pitch wave. As an example, let us add a knob at the first negative peak of the waveform and another knob on the waveform at about 5000 on the horizontal axis. These two knobs appear in Figure 7.32. Select the line_fit option. The mouse cursor changes to a cross hair on the waveform. Press the left mouse button and straight line models are drawn between the knobs, as shown in Figure 7.32. Pressing the Undo button can reverse this action. The other line drawing models work in the same manner. Once the user places the desired knobs

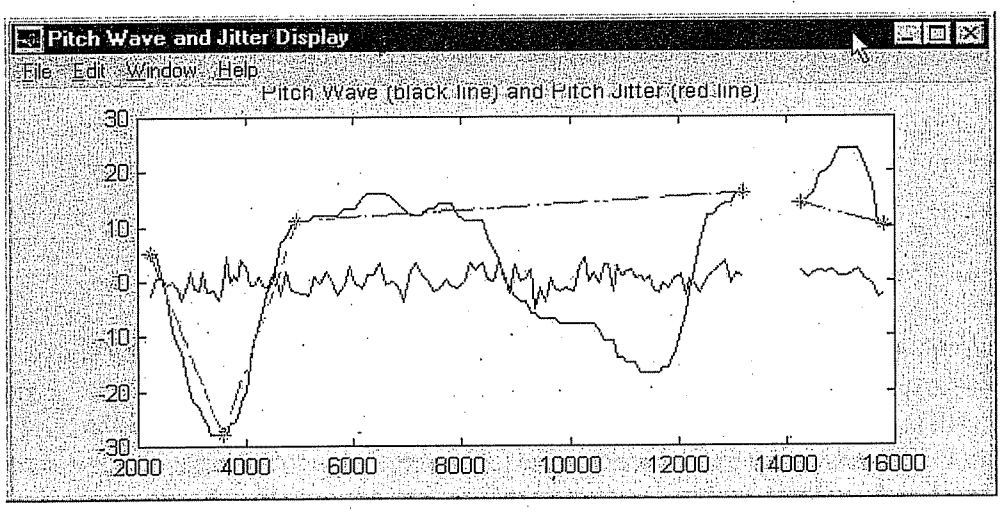

FIGURE 7.32 Pitch wave and jitter window for a simple example.

on the waveform, the various line drawing models can be tried in succession. For example, the straight line fit can be made as described previously. To change the straight line fit to the parabola-fit(2), for example, select the parabola-fit(2) model. The mouse cursor changes to a cross hair. Move the cross hair to a location between knob 1 and knob 2. Click the left mouse and a parabola is fit between these two successive knobs. Repeat the entire process again for the next two successive knobs, until parabola-fit(2) is redrawn as a model fit to the waveform. Of course, an alternate procedure is to undo the line_fit model, place new knobs on the waveform, and select the parabola-fit(2) model.

The parabola-fit models attempt to avoid discontinuities at knobs and also attempt to fit the models to the data. As a consequence, sometimes the fit by a parabola model between the knobs may be an upward or downward parabola, or even a straight line.

A knob can be removed (deleted) by selecting the delete knob option. The mouse cursor changes to a cross hair on the figure. Move the cross hair to the desired knob and click the left mouse button. The knob will be circled. The mouse cursor returns. Next click the Undo button and the knob will be deleted. A knob can be moved by pressing the Select Knob button. The mouse cursor changes to a cross hair on the figure. Select the desired knob to be moved with the cross hair and press the left mouse button. The mouse cursor returns. Press the Up or Down button repeatedly to move the knob accordingly. Select the desired line drawing model. Move the cross hair to a location left of the moved knob and press the left mouse button. Note that the drawing retains a memory of the location of the previous knob location for the segment to the right of the knob. Repeat the previous operations for the segment to the right of the knob. A line model is drawn for this segment. Remember that at any time, the Undo button can be pressed to erase the line drawing models. Only press the Apply button after you are sure that the desired model is the one you want. If you should do this and find you want to cancel the operation, then you must start over by loading again the desired x.mat file, for example file m11_poly_lp_anal.mat. Press the Return button to close these three windows and return to the main modification function window.

7.10 GAIN CONTOUR

The three windows for modifying the gain contour are similar to those shown for the Pitch contour. Figure 7.33 is the action window, while Figure 7.34 shows the gain envelope and the gain perturbation. Figure 7.35 depicts the gain contour. Note that the voiced (V) and unvoiced (U) regions are labeled on Figures 7.34 and 7.35. The gain envelope and perturbation are altered separately as follows. The perturbation is controlled by a slider. Press the left side of the slider to decrease the value, or the right side to increase the value. The procedures for modifying the gain envelope are similar to those for modifying the pitch wave as discussed.

A knob can be added or deleted as described for the Pitch contour. And the Gain contour can be modeled with the same line drawing options.

The label grossly option is the default and provides a plot of the data with two knobs at either end. With this option, the user can add knobs as desired and select the desired line drawing option to model the gain contour. Only voiced speech should be modeled. The label finely option shows the actual data values on the gain contour waveform as + signs. With this option the user can zoom in and out on the data. Knobs cannot be added or deleted with this option. A knob can be selected and moved, however. Line drawing options can be selected.

To cancel a modification, click the Undo button. Press the Apply button when all modifications are completed. Press the Return button to close these three windows and return to the main Modification function window.

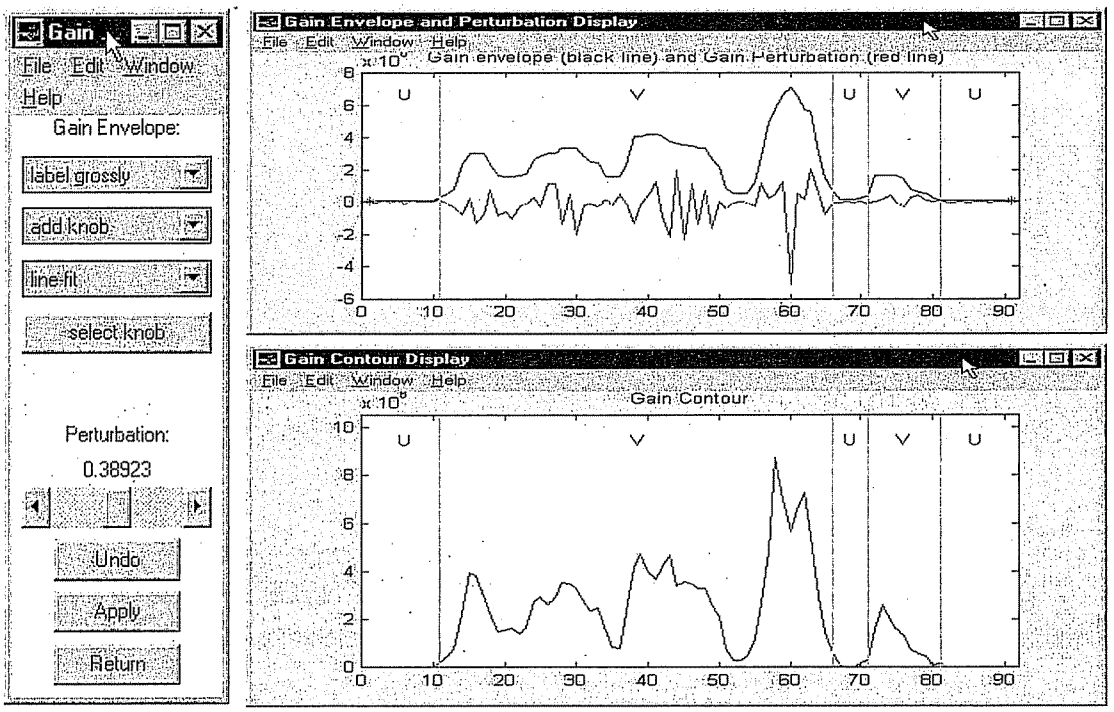

FIGURE 7.33–7.35 Gain modification window (left). Gain envelope and perturbation window (upper right). Gain contour window (lower right).

7.11 GLOTTAL WAVEFORM

This modification option is not available for the polynomial source model. It can be used only for the LF source modeling. Figures 7.36 and 7.37 show the windows for modifying the differentiated LF glottal flow model for the file m11_lf_for_anal.mat. Recall that this file used the LF source model and the formant vocal tract model. Figure 7.36 shows the function buttons and sliders for altering the LF timing parameters. The user can also select the frame of interest using either the < down or up > buttons or the slider. Figure 7.37 displays the

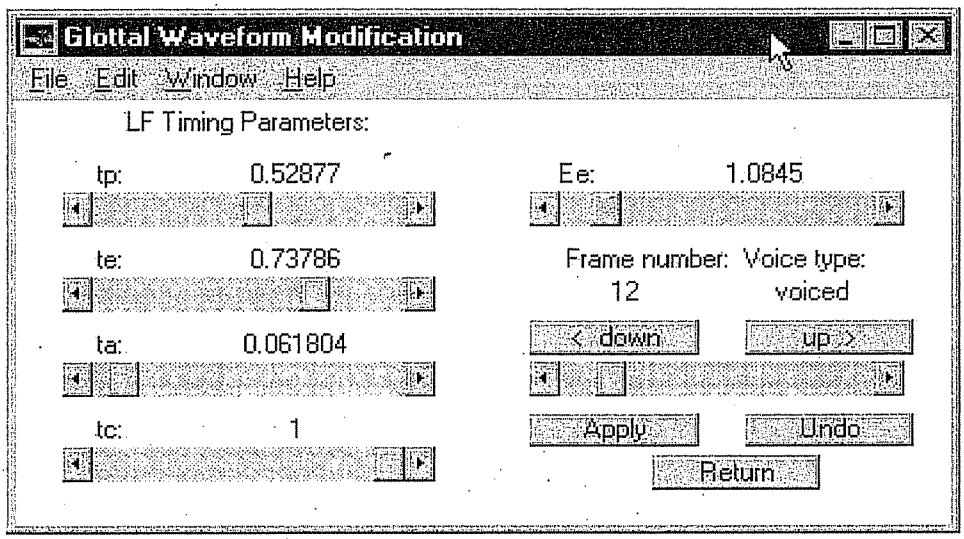

FIGURE 7.36 Modification window for the differentiated LF glottal waveform model.

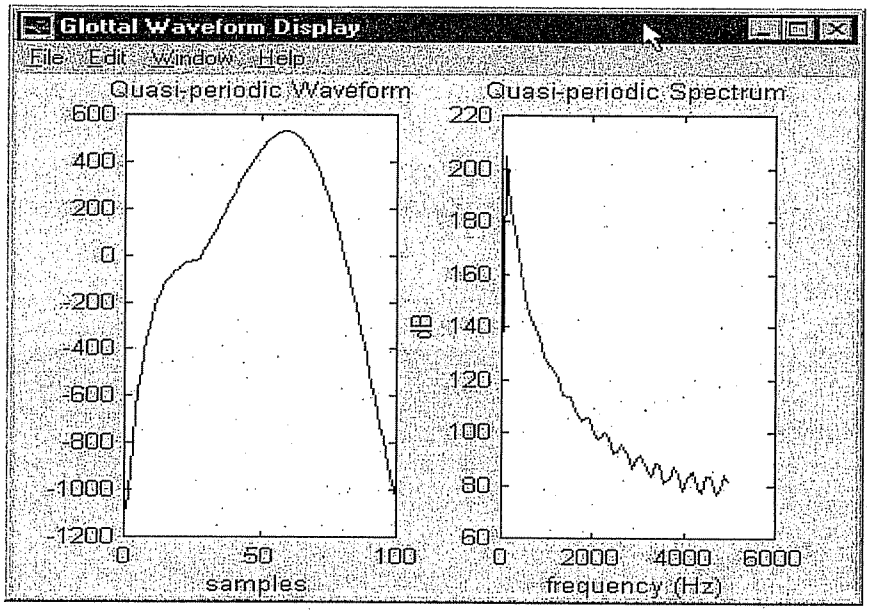

FIGURE 7.37 Differentiated LF glottal waveform model and its spectrum.

LF differentiated glottal waveform model and its spectrum for the selected frame. The user should be careful about setting the LF timing parameters. An error message occurs if the setting is not a valid setting for a model parameter, for example, if T_p is larger than T_e.

7.12 SPECTROGRAM (FORMANTS)

This modification function is not available for use with the LP vocal tract model. It can be used only with the formant vocal tract model. Figures 7.38 and 7.39 show the windows for

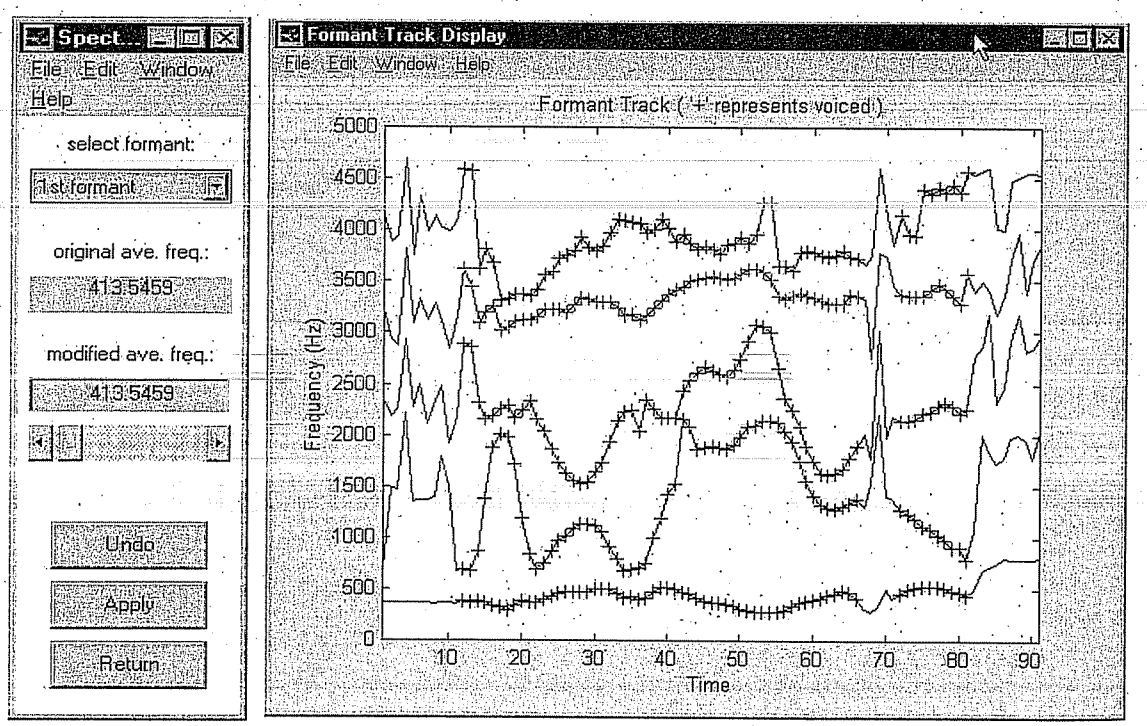

FIGURE 7.38–7.39 Formant modification window (left). Formant tracks (right).

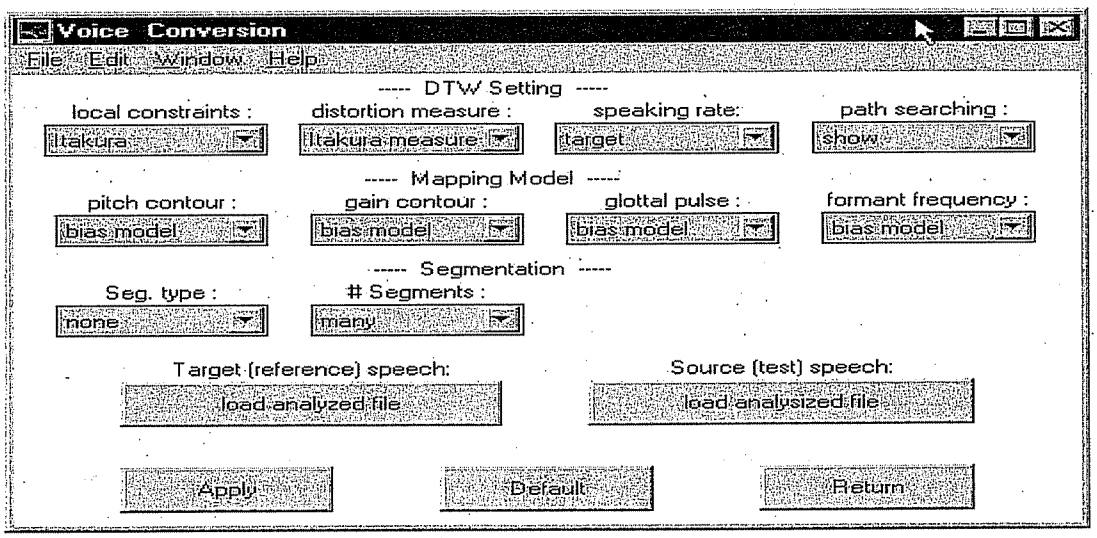

FIGURE 7.40 Voice conversion window.

modifying the spectrogram (formants) of the synthesized speech. Figure 7.38 provides the control buttons for shifting a formant track. Only the average value of the formant track can be modified. As a change is being made the new formant track is plotted along with the old track to facilitate the assessment of the change. The Undo button cancels any changes. After making the desired modifications, the user clicks the Apply button followed by a press of the Return button.

7.13 VOICE CONVERSION

Figure 7.40 shows the various options for voice conversion. The first row of four buttons is for dynamic time warping (DTW). These options let the user set the local constraints and the distortion measures, select the speaking rate as the source or the target, and display the search path. The second row of four buttons provide four models (bias, linear, copy, retain) for mapping the four acoustic features (pitch contour, gain contour, glottal pulse, formant frequency). The two buttons in the third row let the user set the type of segmentation for the speech file. We discuss the segmentation option later.

The fourth row of two buttons load the Target and Source speech files, which must be obtained via analysis and stored as mat files. The Default button resets the option choices to the default values. Once the various options are selected, and the necessary files loaded, press the Apply button to start the voice conversion process. The Return button closes the window and returns to the Modification window. As an example, suppose we select the default values (no segmentation) in the Voice Conversion window and load the target speech file, m11_poly_lp_anal.mat, and the source speech file, m21_poly_lp_anal.mat, and press Apply. As the voice conversion process calculations are underway, various windows are displayed related to the DTW calculations. These displays are only indicative of the results and serve only as indications that calculations are being made. Upon completion to the voice conversion process, a message appears in the MATLAB Command window informing the user that the synthesizer can be run. Close the Voice Conversion window and select Synthesis in the Main window. It is not necessary to load files for synthesis, since the files for voice conversion are in memory. Press the Synthesis button to synthesize the converted speech. Figure 7.41 shows the results.

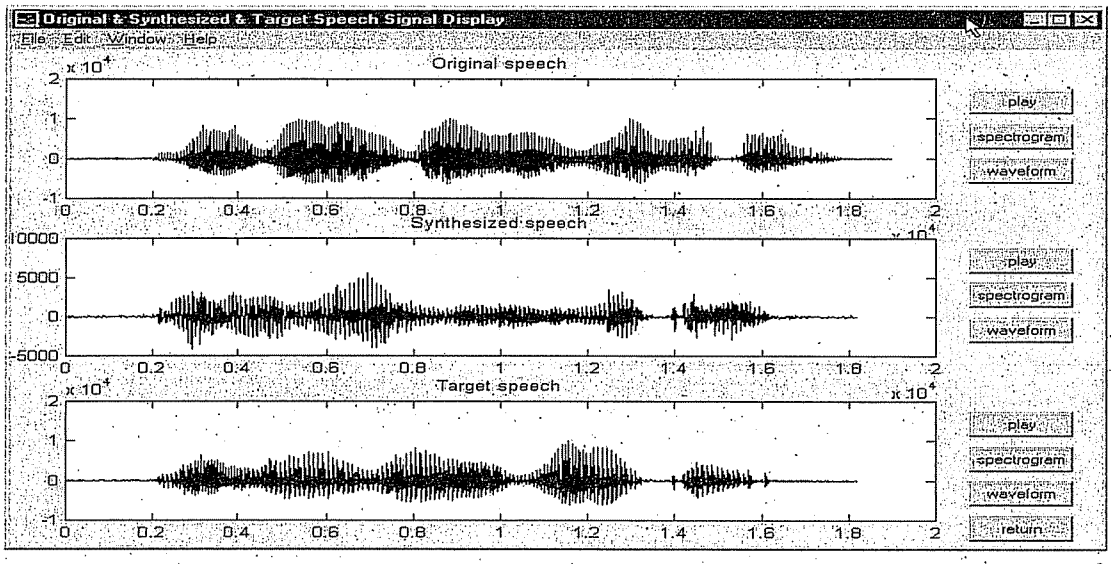

FIGURE 7.41 Synthesis window for voice converted speech waveforms.

Note that the target and source mat files must be for the same vocal tract model, that is, both must be obtained using either the LP or formant vocal tract model, but one file cannot be for the LP model and the other file for the formant model.

The results displayed in Figures 7.41 and 7.42 are for the original (m21.dat) and the target speech (m11.dat). The synthesized speech is the converted speech, that is, the data calculated using voice conversion. The synthesized results are obtained by converting

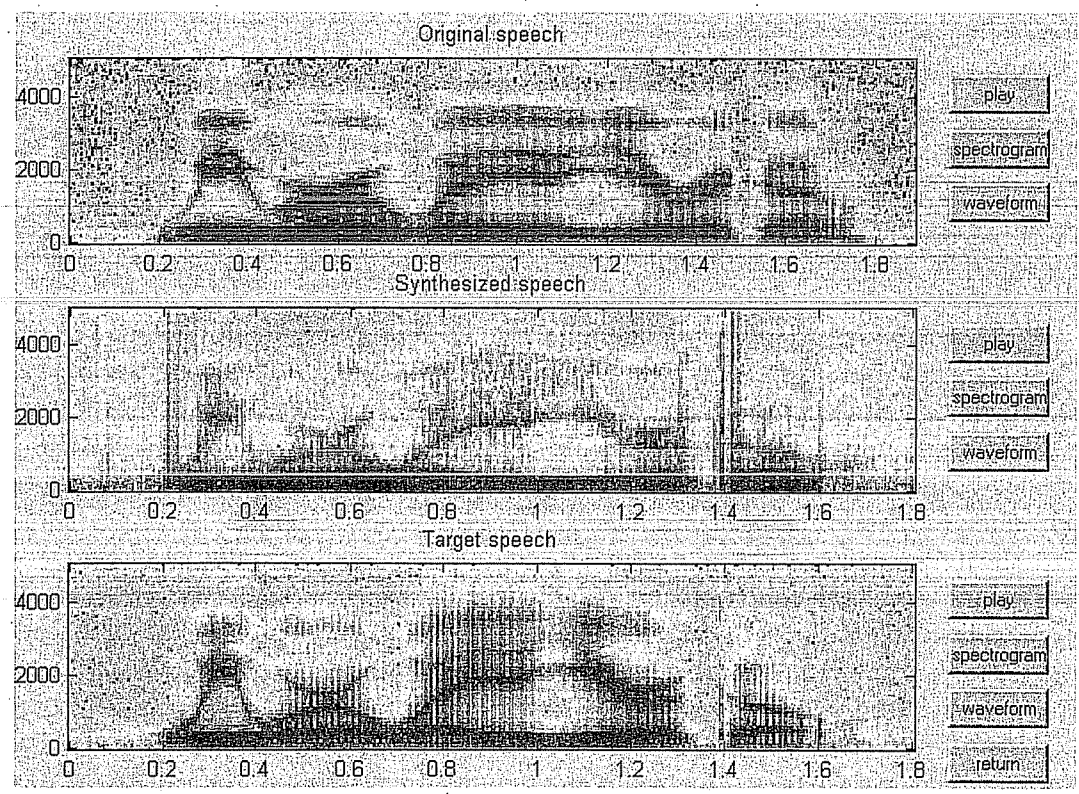

FIGURE 7.42 Synthesis window for voice converted speech spectrograms.

m21_poly_lp_anal.mat to m11_poly_lp_anal.mat. Figure 7.41 displays the speech data for the original and target, i.e., the m21.dat and m11.dat files, while Figure 7.42 shows the spectrograms for the same data. Another comparison is to show the synthesized files for all three, that is, original (source), converted, and target. We show how to do this below.

7.14 VOICE CONVERSION—SEGMENTATION

The purpose of the segmentation option is to improve the voice conversion process. To convert the parameters of one voice to that of another for a complete sentence can be a difficult task. However, doing voice conversion on a segment-by-segment basis can sometimes give improved results.

Segmentation is performed over the target speech using a simple normalized measure of spectral change given by:

$$\frac{\sum_{\omega}(|S_1(\omega)|-|S_2(\omega)|)^2}{\sum_{\omega}|S_1(\omega)|^2} \geq \text{threshold} \qquad (7.14.1)$$

S_1 and S_2 are taken from two consecutive frames. If this value exceeds the threshold, then a new segment is specified. The segmentation buttons apply only if the speaking rate is set to the target. The Segmentation type choices include without segmentation (none), with segmentation (automatic), and with manual segmentation (manual). If the Segmentation option is selected, then the user specifies the number of segments via the # Segments pull down menu. The choices are many, medium, and few. The choice many corresponds to a threshold of 0.2, medium corresponds to a threshold of 0.4, and few to 1. A segmentation of one frame is not allowed. Furthermore, when the option many segments is selected, the analysis can calculate numerous one-frame segments. Thus, sometimes it is better to select a medium number of segments.

The segmentation option must be used as follows. The speaking rate must be set to target. The target and source files must be for the formant vocal tract model. The LP vocal tract model is not available for segmentation.

Once the user decides to segment the data, there are two options: automatic and manual. For each of these options, the threshold value can be set. This is done as described, by selecting the number of segments to be many, medium, or few. If the user selects, for example, automatic segmentation and medium number of segments, loads the desired data files, and presses the Apply button, then there are no other options. Upon completion of the calculations, a window shows the data with superimposed segments. The user can save the results and synthesize.

If the user selects, for example, manual segmentation and medium number of segments, loads the two files; m21_lf_for_anal.mat (target) and m11_lf_for_anal.mat (source), and presses the Apply button, then after the calculations are completed the window shown in Figure 7.43 appears. In this window the user can modify the segmentation results in a manner similar to that described for GCI and formant tract modification. The user can zoom in and out and add and delete a segment. To add a segment mark (or boundary), press the Add button, move the cross hair to the desired location and press the left mouse button. A new segment marker is placed at this location and the data between the new segment marker and the one to the left of this marker is played. Another option is to move the cross hair to the desired location and click the right mouse button. The data between the location of the

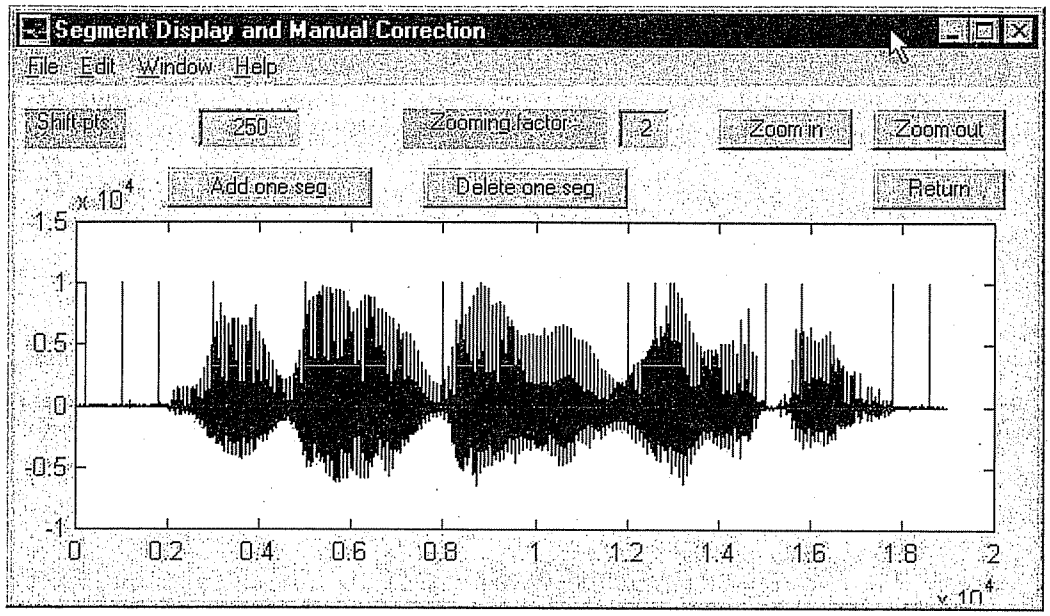

FIGURE 7.43 Manual segmentation window.

cross hair and the segment marker to the left of the cross hair is played. A new segment is not added until the left mouse button is clicked.

To delete a segment marker follow a similar procedure, clicking the left mouse button when the cross hair is placed on the segment marker to be deleted. When the user adds or deletes a segment marker the Apply button appears. Press this button to make the desired changes permanent. If the user presses the Return button without pressing the Apply button, then all changes are discarded. Pressing the Return button automatically initiates the continuation of the execution of the program.

The segmentation option does pitch conversion on a segment-by-segment basis by placing a restriction on the placement of the GCIs at segment boundaries to avoid discontinuities. The gain contour conversion process is smoothed with a median filter of length 7. The glottal source conversion is the same as that without segmentation, except the conversion is done on a segment-by-segment basis. The vocal tract conversion is performed only for the formant vocal tract model and is done in the same manner as that for no segmentation. In general, the formant bandwidths are not converted. The user can select either source model.

7.15 ALTERNATE LOAD, DISPLAY, AND PLAY OPTION

The m-file called speech_display.m in the display_speech folder within the VOCOS folder can be used to load, play, and compare four speech dat files. Change directory to the display_speech folder and type speech_display in the MATLAB Command window to bring up the window in Figure 7.44. The user can load speech or synthesized speech data (dat) files. These files can be played, and the waveforms or spectrograms can be compared. Figure 7.44 shows the files f11.dat, f21.dat, m11.dat, and m21.dat.

To use the speech_display.m file, first load the desired number of data files, that is, 1 to 4 files. A message appears in the MATLAB Command window reminding the user to strike any key on the key board to continue. Upon doing so, the window shown in Figure 7.45 appears. The user can select one of the various buttons, play or spectrogram, for each of the

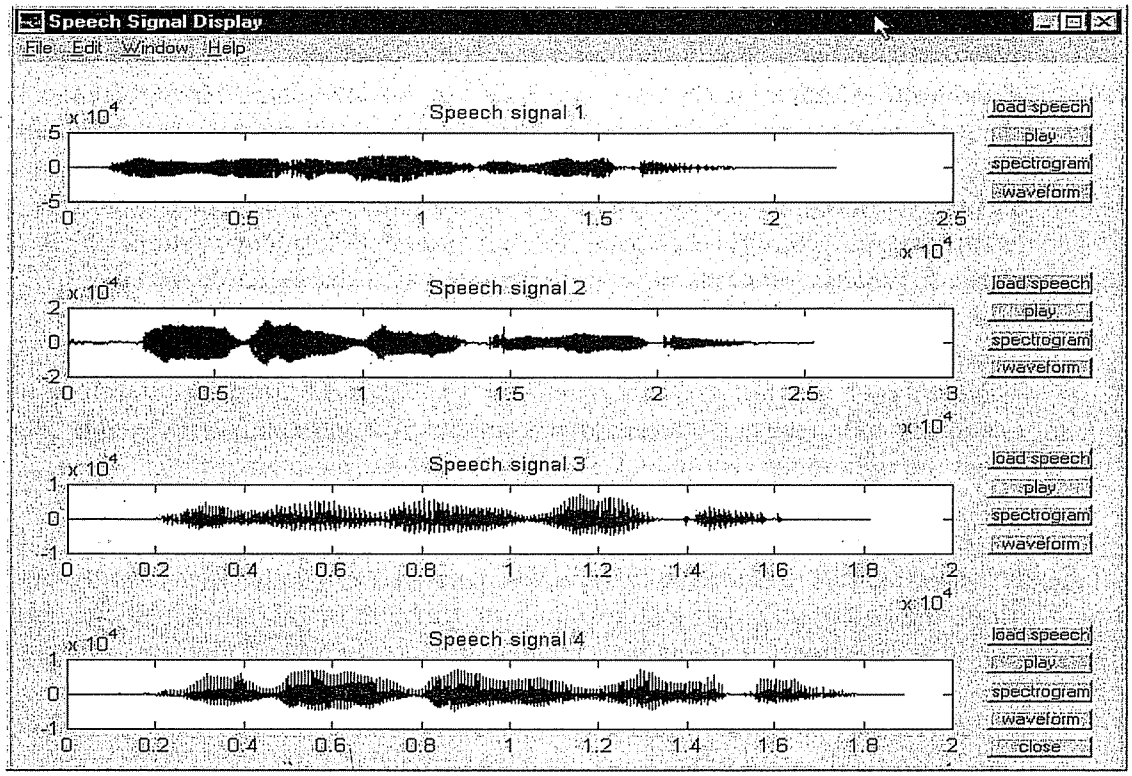

FIGURE 7.44 Speech signal display window to compare up to four speech signals.

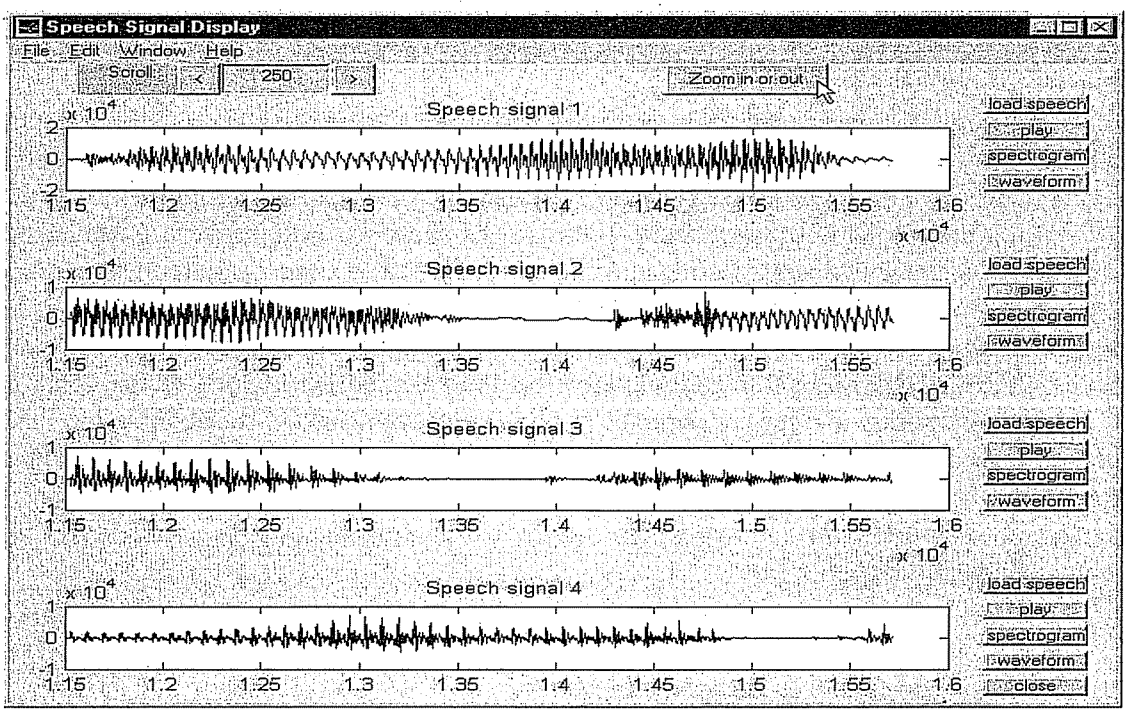

FIGURE 7.45 Speech signal display after loading the data files and pressing any key.

speech signals. The waveform button redisplays the speech signal if the spectrogram is plotted. Note that a Zoom-in or -out button is available to zoom the waveform data display. Pressing this button changes the mouse cursor to a cross hair. Move the cross hair to the desired initial x-axis data location to zoom in. Press the left mouse button once. Move the cross hair to the desired end of the zoom in x-axis segment. Press the left mouse button once again. This zooms in one level on all the loaded data files at the same time. The zoomed-in waveforms are plotted in each panel. Repeat these steps to zoom in another level. To stop the zoom-in process, press the right mouse button twice slowly while the cross hair is visible. The cross hair disappears and the standard mouse cursor reappears. To zoom out one level, press the Zoom-in or -out button. The cross hair reappears. Press the left mouse button, followed by a press of the right mouse button. The data are zoomed out one level. Each repetition of this sequence zooms out another level, until the original signal level is reached. Press the right mouse button twice slowly to exit the zoom out option.

The scroll option is to be used after zooming in on the data at least one level. The scroll option becomes available automatically after zooming in. The default scroll value is 250 data points. This value can be changed by typing in a new value in the scroll window panel. Press the left (right) mouse button to scroll the data to the left (right). You can continue scrolling in either direction until the end of the data record. The scroll option can be used at any zoom-in level, as long as the zoom-in level is at least one.

The play option is designed to play the data displayed. Thus, if the data are zoomed in, the data played are the data shown in the panel. Similar remarks apply to the spectrogram option.

The Close button closes out the Speech Signal Display window.

7.16 SUMMARY

There are three major features to the voice modification system: (1) the analysis and parameterization of acoustic features; (2) parameter visualization; and (3) windows for user interaction to modify parameters. There are four aspects of the software system: analysis, correction, modification, and synthesis. The analysis phase extracts the acoustic features from the speech signal using specific algorithms. The acoustic features are represented by sets of parameters, which are displayed in graphs. The correction phase allows the user to change parameter values. The modification phase is similar to the correction phase, but is more extensive, allowing parameter values to be altered (modified). One purpose for the modification phase is to alter the parameters to convert one speaker's voice to sound like that of another speaker's voice, that is, voice conversion. Finally, the synthesis phase allows the synthesis of speech using the measured parameters. Synthesis can be used for voice conversion or more simply to synthesize a speech file using the parameter file obtained using the analysis phase.

7.17 MISCELLANEOUS MATTERS

Sometimes the user will make an error in the use of the software, perhaps making an incorrect selection, or clicking the wrong option or the wrong mouse button. When this happens a software variable may be set incorrectly, leading to an incorrect software calculation or data plot. A beginning user may not notice such errors at first. However, as the user becomes more skilled such errors are apparent. The best way to reset the software is to quit the

analysis program and restart, loading the data file and selecting the desired options again. This is necessary only occasionally.

Do not move and click the mouse button while the software is making calculations, since this can cause the data to be plotted in an incorrect window.

PROBLEMS

The following set of problems is to be completed using the default parameter settings in the analysis specification window. This means that the polynomial excitation model is used and the vocal tract filter is LP. Once Problems 7.1 through 7.6 (or a subset) have been completed, perform the same task using the polynomial model and the formant vocal tract model. Repeat using the LF and the LP models. Repeat using the LF and formant models

7.1 Use the software in this chapter to convert the sentence, "We were away a year ago," for a male voice in file m0125s.dat to sound like that of a female voice for the same sentence in file f0625s.dat.

7.2 Use the software in this chapter to convert the sentence, "We were away a year ago," for a female voice in file f0625s.dat to sound like that of a male voice for the same sentence in file m0125s.dat.

7.3 Repeat Problem 7.1 for the sentence, "Early one morning a man and woman ambled along a one mile lane," using files m0126s.dat and f0626s.dat.

7.4 Repeat Problem 7.2 for the sentence, "Early one morning a man and woman ambled along a one mile lane," using files m0126s.dat and f0626s.dat.

7.5 Repeat Problem 7.1 for the sentence, "Should we chase those cowboys?" using files m0127s.dat and f0627s.dat.

7.6 Repeat Problem 7.2 for the sentence, "Should we chase those cowboys?" using files m0127s.dat and f0627s.dat.

7.7 Analyze the sentence, "We were away a year ago," for the file m0125s.dat. Use the default settings for the polynomial and LP models. Experiment with altering the pitch contour and the jitter contour to create a breathy voice, a hoarse voice, a falsetto voice, and a vocal fry voice. It may be necessary to alter the gain contour as well.

7.8 Repeat Problem 7.7 for the sentence, "Should we chase those cowboys?" for file m0127.dat.

7.9 Analyze the sentence, "We were away a year ago," for the file m0125s.dat. Use the default settings for the polynomial and LP models. Experiment with altering the gain contour to determine the influence of this parameter on the quality of the voice.

7.10 Repeat Problem 7.9 for the sentence, "Should we chase those cowboys?" for file m0127.dat.

7.11 Analyze the sentence, "We were away a year ago," for the file m0125s.dat. Use the default settings for the polynomial and LP models. Experiment with altering the formant contours to determine the influence of this parameter on both the quality of the voice and the intelligibility of the synthesized speech.

7.12 Repeat Problem 7.11 for the sentence, "Should we chase those cowboys?" for file m0127.dat.

7.13 Return to Problem 7.1. Experiment with the various voice conversion parameters for dynamic time warping (DTW) to determine the effect these parameters have on voice quality and speech intelligibility.

7.14 Repeat Problem 7.13 except return to Problem 7.5.

7.15 Return to Problem 7.1. Experiment with the segmentation process to determine the effect this process has on voice quality and speech intelligibility.

7.16 Repeat Problem 7.15 except return to Problem 7.5.

7.17 Analyze the sentence, "We were away a year ago," for the file m0125s.dat. Use the default settings for the polynomial and LP models. Experiment with altering the formant contours to determine if you can create a voice with an accent or dialect, such as a Spanish accent (or one of your choice). It might be helpful to analyze the same sentence spoken by a person with such an accent.

7.18 Analyze the sentence, "We were away a year ago," for the file m0125s.dat. Use the default settings for the polynomial and LP models. Experiment with inserting pauses at various locations within the sentence to create a new effect in the synthesized speech.