GUI-STRAIGHT: Getting started (for ver.23)

Hideki Kawahara
ATR Human Information Processing Research Laboratories
Faculty of Systems Engineering, Wakayama University
CREST[*]

draft 4:32 P.M., March 22, 1998

Contents

[*] CREST stands for Core Research for Evolutional Science and Technology of Japan Science and Technology Corporation (JST).

1 Introduction

STRAIGHT-suite[1] is a set of procedures for analyzing, modifying, and synthesizing speech-like sounds. The recent introduction of GUI-STRAIGHT[2] has made it extremely easy to use STRAIGHT. This document provides a step-by-step introduction of GUI-STRAIGHT.

2 System requirements

The current version of GUI-STRAIGHT operates on the following platforms:

2.1 Software requirement

MATLAB version 5.0 (or later) and signal processing toolbox are recommended. No other toolboxes are necessary to run GUI-STRAIGHT. Some component procedures of STRAIGHT may be functional on older versions of MATLAB; however we strongly recommend using version 5.0 or later.

For platforms other than a Macintosh system, it is also recommended to have an audio input function able to be called from inside MATLAB. It is still possible to use a separate software package for audio input/output. GUI-STRAIGHT supports AIFF (.aiff) and WAVE (.wav) file formats as well as the usual plain binary (16-bit) format.

2.2 Hardware requirements

- **[Macintosh] 68k Macintosh, Power Macintosh, or Power Book.**
- **[Windows95] PC with a Pentium, MMX-Pentium, or Pentium II processor.**
- **[UNIX] SUN, SGI, or HP.**
- **[Linux]**

2.2.1 Machine power

Using machines slower than a Sparc 2 is painfully slow. Earlier versions required about 1Gflops to process a short (1 second long: sampled at 16-bit, 24kHz) utterance.

2.2.2 Memory requirements

The larger the better. :-)
At times, 64MB will not be enough.

3 Tour

This section provides a step-by-step introduction on how to use GUI-STRAIGHT. In the following examples, the verbatim font is used to represent each system prompt and the user commands. ‘‘>>’’ at the beginning of each line is the MATLAB system’s prompt. The figures in this document have basically been captured on a Macintosh.

3.1 Search path

Please start MATLAB on your system. Set the search path to include the STRAIGHT directory. On a UNIX machine, the following command adds the path.

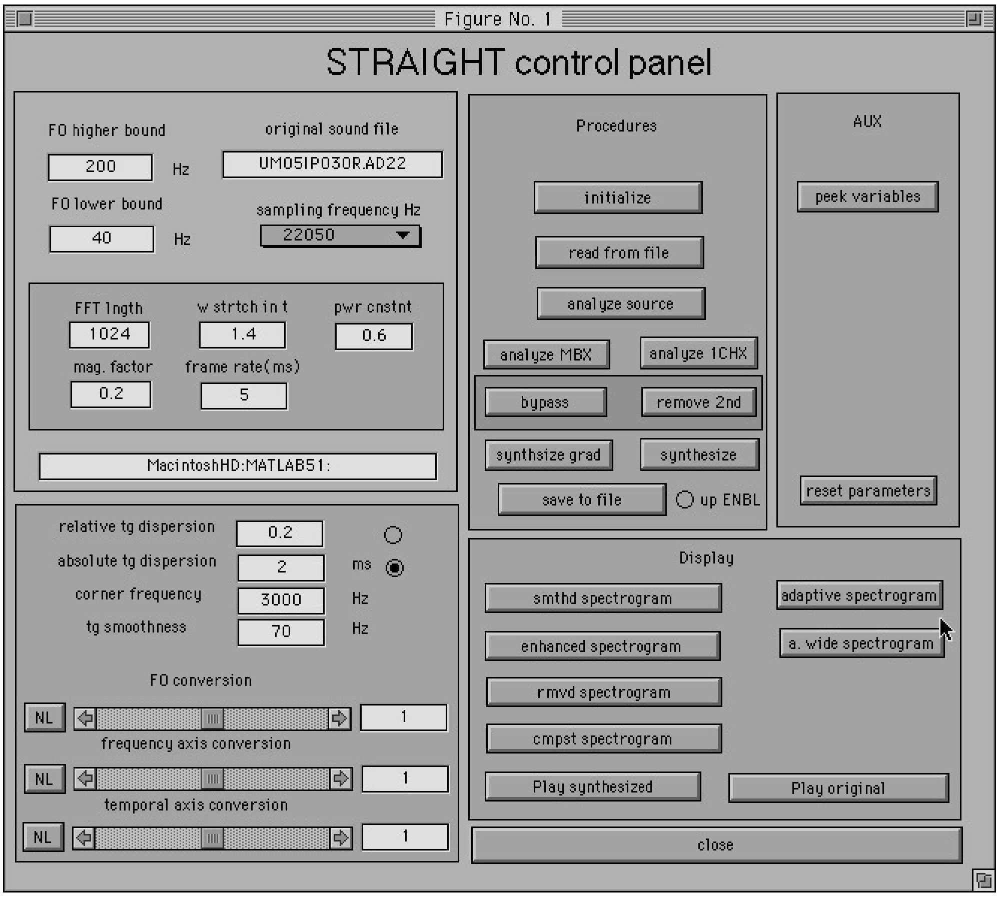
```
>>          path(path,'/usr/people/kawahara/matlab/STRAIGHTV21');
```

In this example, STRAIGHT programs are located in the following directory:

```
/usr/people/kawahara/matlab/STRAIGHTV21/.
```

[1] STRAIGHT stands for Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrogram. It consists of three major procedures STRAIGHT-core, TEMPO, and SPIKES. Please refer to the references listed at the end of this document.

[2] GUI stands for Graphical User Interface.



F0 higher bound: 200 Hz
original sound file: UM05IP030R.AD22
F0 lower bound: 40 Hz
sampling frequency Hz: 22050

FFT lngth: 1024
w strtch in t: 1.4
pwr cnstnt: 0.6
mag. factor: 0.2
frame rate(ms): 5

MacintoshHD:MATLAB51:

relative tg dispersion: 0.2 ○
absolute tg dispersion: 2 ms ●
corner frequency: 3000 Hz
tg smoothness: 70 Hz

F0 conversion: 1
frequency axis conversion: 1
temporal axis conversion: 1

Figure 1: STRAIGHT control panel.

3.2 STRAIGHT control panel

Now, please type the command

to start the system.

The control panel (Figure 1) will be shown in a couple of seconds. Those buttons that are relevant for the current context are made active; inappropriate buttons are disabled. The first button to be clicked is the ‘read from file’ button.

The following sub-subsections briefly introduce which panel is which.

3.2.1 Procedure panel

The top center panel is for general procedures. The usual way of using STRAIGHT-suite is to click buttons from top to bottom. This represents the standard ordering of constituent procedures for manipulating speech.

Figure 2 illustrates four possible paths for manipulating sounds. As shown in the figure, two lines of procedures are currently supported. The paths via ‘analyze 1CHX’ use a couple of single vectors to carry V/UV (voiced/unvoiced) information. The other paths via ‘analyze MBX’ use a continuous-valued V/UV map to represent the periodicity in each time-frequency region.

As shown in the figures, there are two alternatives for the path via ‘analyze 1CHX’ and the path vis ‘analyze MBX’. One is to use the analysis result directly, in manipulation and re-synthesis. The other removes the spectral second-order structure[3] before any manipulation or re-synthesis takes place.

3.2.2 Display panel

This panel has a collection of buttons to display information about the speech sample under inspection. The spectrograms in each step of the STRAIGHT-core procedure are accessible. The panel also provides audio monitoring of the signal.

[3] This concept is new and is not well established. A brief description and exemplar will be introduced in a later section.

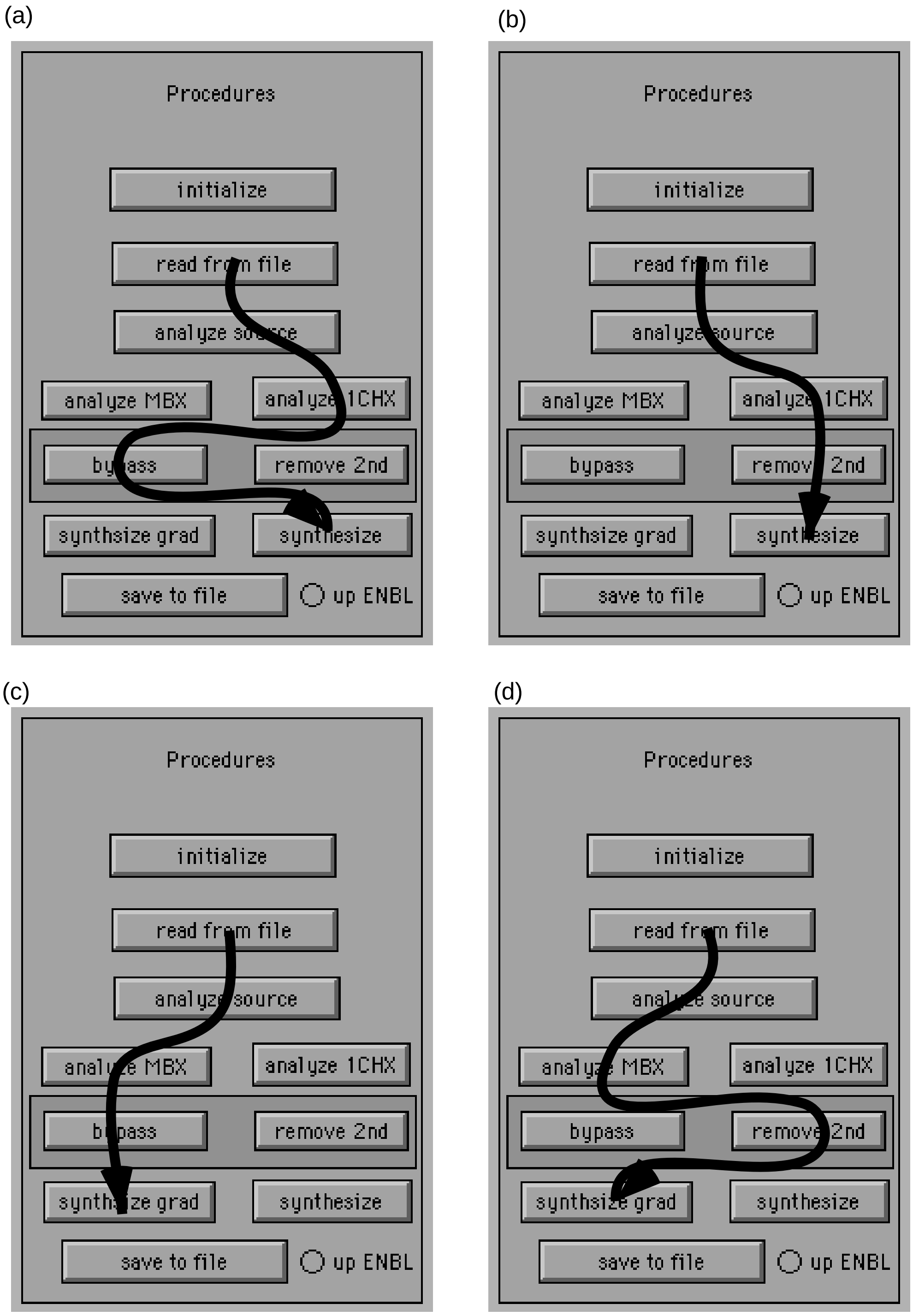


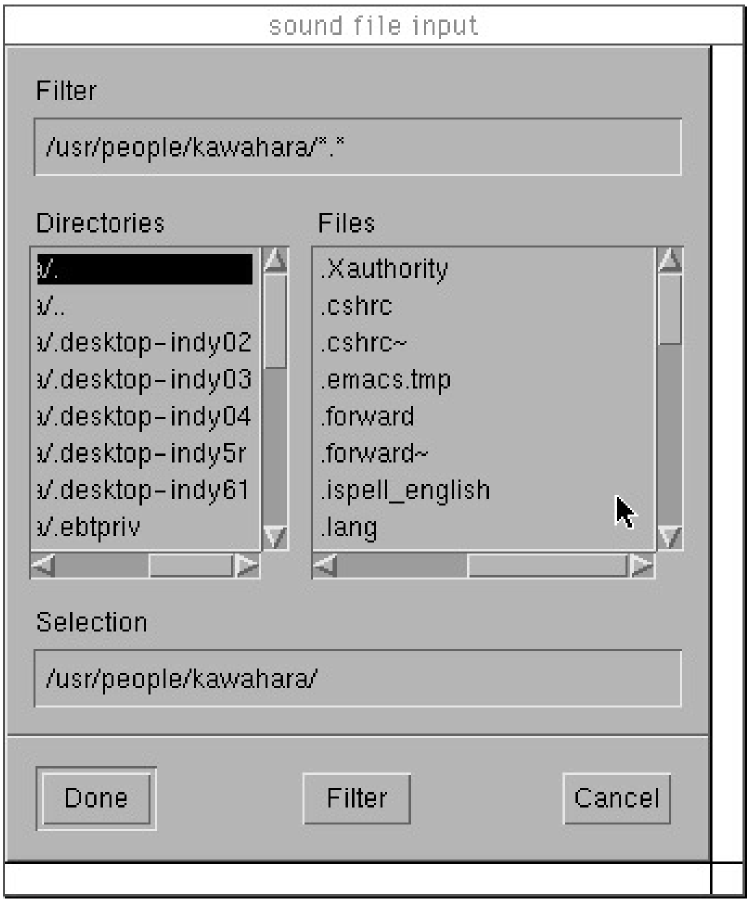
Figure 2: Alternative paths for manipulating sounds.

3.2.3 AUX panel

Miscellaneous functions are located on this panel.

3.2.4 Analysis parameter panel

Parameters mainly used in the analysis stage are accessible from this panel and are controllable using the ‘edit’ and ‘menu selection’ primitives of MATLAB.



/usr/people/kawahara/*.*

/.
/..
/.desktop-indy02
/.desktop-indy03
/.desktop-indy04
/.desktop-indy5r
/.desktop-indy61
/.ebtpriv

.Xauthority
.cshrc
.cshrc~
.emacs.tmp
.forward
.forward~
.ispell_english
.lang

/usr/people/kawahara/

Figure 3: File input user interface for UNIX.

3.2.5 Manipulation and synthesis panel

Parameters used in the synthesis stage are accessible from this panel and are controllable using the ‘edit’, ‘slider’, and ‘radio button’ primitives of MATLAB. Any nonlinear arbitrary mapping of an original and the synthesis parameters can be controlled using a direct manipulation controller.

3.3 Reading a speech file

The next step is to read the speech file to be manipulated. By clicking the ‘read from file’ button, the user interface for file input will appear. The ordinary dialogue file interface pops up for the Macintosh. Figure 3 shows the graphical file interface for a UNIX machine.

The file input routine accepts the WAVE format, the AIFF (and AIFF-C) format, and the plain 16-bit linear binary format. WAVE or AIFF processing is invoked based on the file extension. WAVE format processing is applied to files having ‘.wav’ as the extension. AIFF (and AIFF-C) format processing is applied to files having ‘.aiff’ as the extension. Files with illegal formats are rejected. Files with an unknown extension are assumed as plain binary format files.

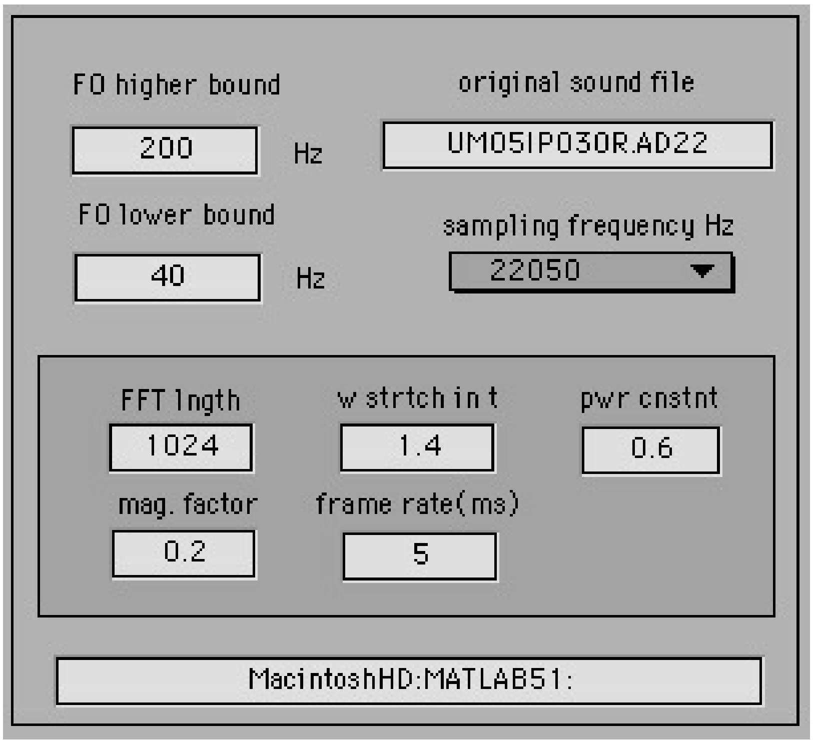
After a file with header information is read, the sampling frequency will be automatically replaced by the value read from the file. For a plain binary file, the user is required to set the sampling frequency manually.

3.4 Analyzing source information (TEMPO)

Since STRAIGHT is a pitch adaptive procedure, it is very important to extract F0 information reliably the first time. Clicking the ‘analyze source’ button invokes source information extraction using TEMPO. **IMPORTANT:** please read the following sub-section before clicking the button.

3.4.1 Notes on analysis parameters

GUI-STRAIGHT provides a means to tweak analysis parameters using the ‘analysis parameter control sub-panel’, which is shown in Figure 4. To speed up the whole process, it is better to modify some parameters from their original values. ‘F0 lower bound’ and ‘F0 higher bound’ define the F0 search range in pitch extraction using TEMPO. For



F0 higher bound 200 Hz

F0 lower bound 40 Hz

original sound file UM05IP030R.AD22

sampling frequency Hz 22050

FFT lngth 1024

w strtch in t 1.4

pwr cnstnt 0.6

mag. factor 0.2

frame rate(ms) 5

MacintoshHD:MATLAB51:

Figure 4: Analysis parameter control sub-panel (top left of the STRAIGHT control panel).

instance, if *a-priori* knowledge about a possible F0 is available, using the information to trim the search range can eliminate unnecessary processing.

By controlling the 'frame rate (ms)[4]', a further speed up is possible. For general use, a 5 ms frame rate is recommended.

Sampling frequency control is implemented as a 'popup menu'. If the file under investigation is either of the WAVE (.wav) or AIFF (.aiff) format, the sampling frequency information in the file header is used to update the sampling frequency on the menu. If the file is ''headerless'', you have to select the proper sampling frequency using this popup menu. The sampling frequencies currently supported are as follows: 8000, 10000, 11025, 12000, 16000, 20000, 22050, 24000, 32000, 44100 and 48000 Hz.

The other parameters (shown on the sub-panel) have the following functions.

- **FFT length** This parameter is automatically set, based on the sampling frequency. The internal FFT uses this length as the frame length. The length is the smallest 2^N (where N is an integer) to cover a 40 ms analysis frame length.

- **w strtch in t** This should be spelled out as ''window stretching factor in the time domain''. This determines η in the definition of a set of compensatory time windows.

- **pwr cnstnt** This should be spelled out as ''power constant''. This determines nonlinear mapping function $g(x)$ for the smoothing operation. Specifically, the nonlinear function of absolute spectral value x has the following form.

  $$g(x) = x^{\alpha} \tag{1}$$

  where α represents the power constant. The default value $\alpha = 0.6$ approximates $g(x)$ for a loudness-preserving smoothing operation.

- **mag. factor** This represents the magnification factor in the time domain. This factor was introduced in my ASJ technical meeting article in July 1997. This time domain operation enhances the sharpness of the spectral peaks. In retrospect, however, this parameter is excessive at least for the current implementation of STRAIGHT-core. Increasing this value does add some 'richness in resonance', but may also introduce an artificial timbre. The default value 0.2 is tentative. Setting 0 for this factor makes the enhancement process do nothing.

[4] The original STRAIGHT's constraint that the rate should be smaller than 1 ms is no longer a constraint. Using a compensatory set of time windows eliminates the need for temporal smoothing.

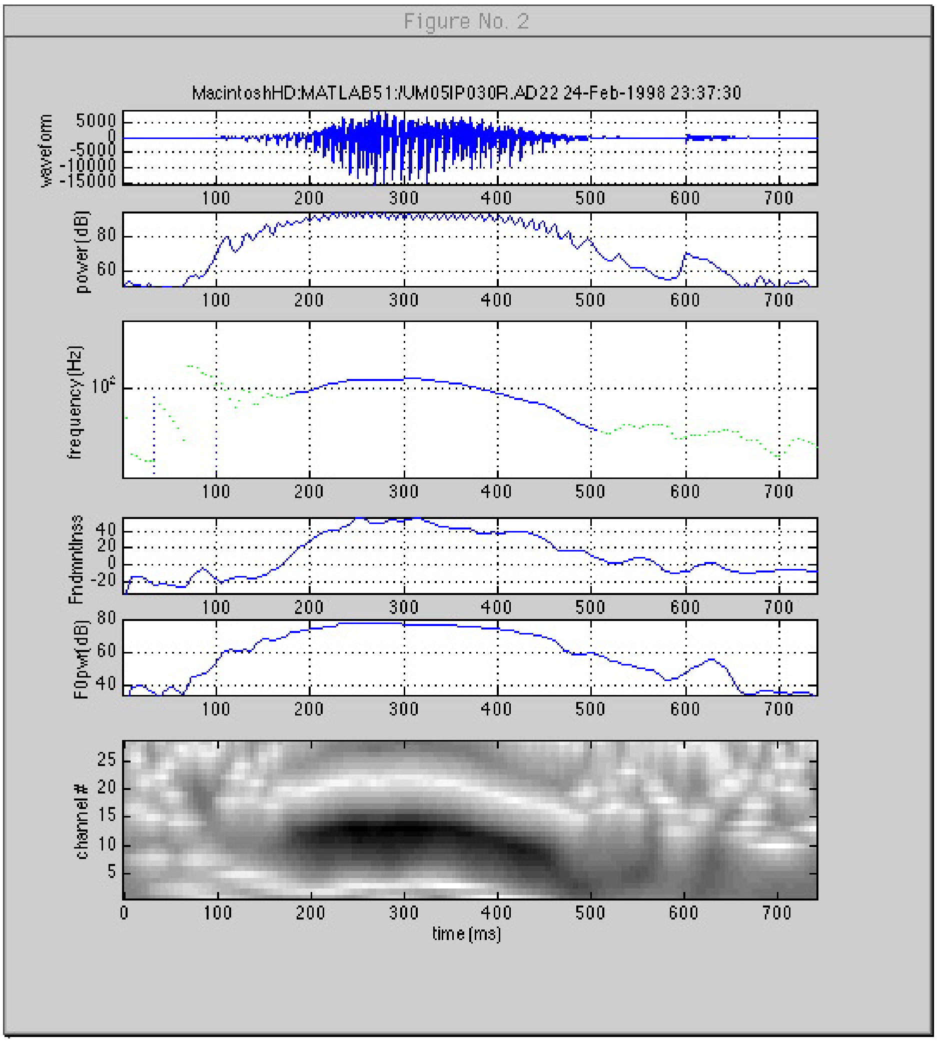


Figure 5: Extracted source information for a male utterance ‘right’. (From top to bottom: waveform, power in the F0 search range, extracted F0, ‘fundamentalness’, F0 power, and ‘fundamentalness’ map.)

3.4.2 Source information display

After the ‘analyze source’ button is clicked, it will take several minutes to calculate the necessary source information for later STRAIGHT processing.

Figure 5 shows the source information display. The top chart displays the input waveform. The horizontal axis (time in ms) and the amplitude axis (unit is LSB) are automatically scaled. The second chart shows the power within the F0 search range. The third plot represents the extracted F0 information. An ad-hoc procedure is used to categorize the F0 trajectory into two parts; blue solid line(s): voiced portion(s) and green dots: unvoiced portions. The ad-hoc procedure does not perform any other functions than this display in the current implementation, because the logic is already obsolete and unreliable. This part will be replaced in a later release.

The fourth plot shows the highest ‘fundamentalness’ at each frame. This value is (roughly speaking) inversely proportional to the log rms F0 errors. The fifth plot is the power of the fundamental component.

The last image is a ‘fundamentalness’ map. The calculated ‘fundamentalness’ of each channel is color coded and mapped onto the time-(log)frequency plane. The vertical axis represents the channel ID for each band-pass filter. The center frequency of channel-1 corresponds[5] to the ‘F0 lower bound’ defined in the analysis sub-panel. The channel interval is set to $2^{1/12}$.

3.5 Extracting a spectral envelope (STRAIGHT-core)

The next step is spectral analysis and envelope extraction. Please click the ‘analyze 1CHX’ button or ‘analyze MBX’ button to start the STRAIGHT-core procedure. After completion of the procedure, the ‘bypass’ and ‘remove

[5] This explanation is not precise. The differentiation operation in the TEMPO procedure introduces some bias to move the effective center frequency upward. This will be revised in a later release.

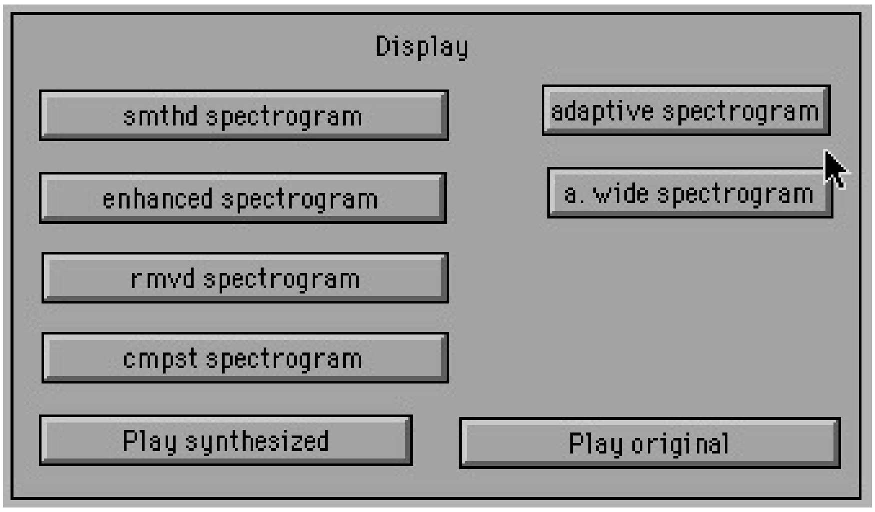


Figure 6: Display control sub-panel.

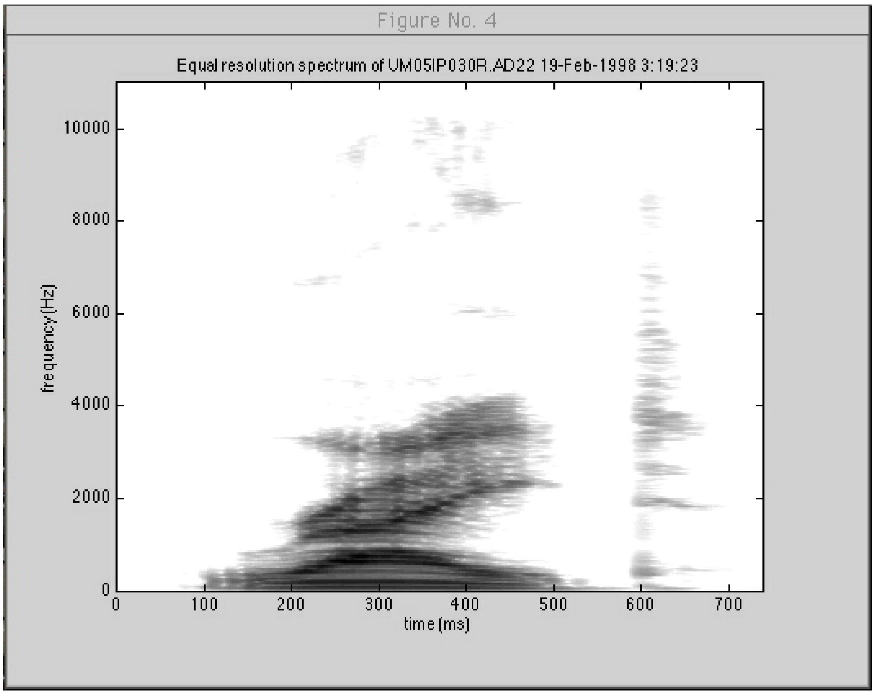


Figure 7: Spectrogram with a compensatory set of adaptive time windows.

2nd' buttons will be made active. Other appropriate buttons on the display sub-panel are also activated.

3.5.1 Case 1: 'analyze 1CHX'

Figure 6 shows the display control sub-panel. After completion of the analysis procedure, the 'adaptive spectrogram', 'smthd spectrogram', and 'enhanced spectrogram' buttons are made active. The 'play original' button is already active.

Figure 7 shows an adaptive spectrogram based on the F0 information of a TEMPO output. To get this image, please click the 'adaptive spectrogram' button. Since the resolution in the frequency domain is set higher than that in the time domain, the harmonic structure is resolved. It should also be noted that even without the temporal smoothing in the original STRAIGHT, the spectral variations in the time domain are very small.

Note that the most salient horizontal structure which resembles the harmonic structure is made from harmonic pairs. In other words, a regular level alternation is observed between adjacent harmonic components. This is strange, but consistent with the observation by Honda in the 1980s in his speech coding papers. This regular spectral structure implies that there is an additional excitation at the precise center position between adjacent primary excitation pulses.

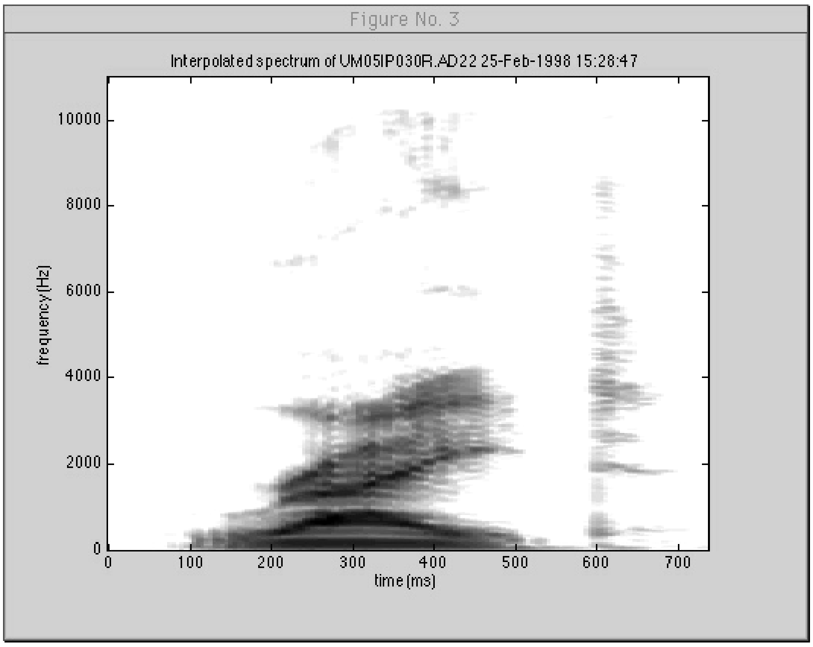


Figure 8: Smoothed and optimally recovered spectrogram. STRAIGHT-core smoothing in the time domain with an optimal smoothing function is employed.

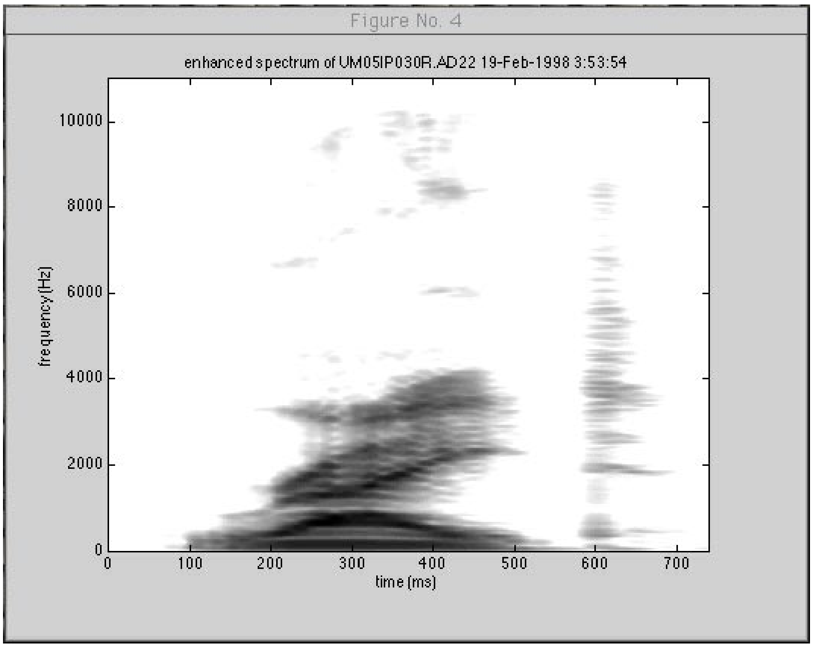


Figure 9: Enhanced spectrogram made from the smoothed spectrogram.

This regular structure introduces degradation when parameters are manipulated before re-synthesis.

Figure 8 shows an optimally smoothed spectrogram using the STRAIGHT-core procedure with smoothing optimization. To get this image, please click the ‘smthd spectrogram’ button.

Figure 9 shows an enhanced spectrogram using time domain processing. To get this image, please click the ‘enhanced spectrogram’ button.

Both of the above spectrograms illustrate that the secondary structure described in the previous paragraph remains intact (or worse, enhanced) in the smoothed spectrograms, where interference (represented as the harmonic structure) is effectively eliminated. This type of secondary structure is often observed in a male utterance. In a few cases, this structure is also found in a female utterance.

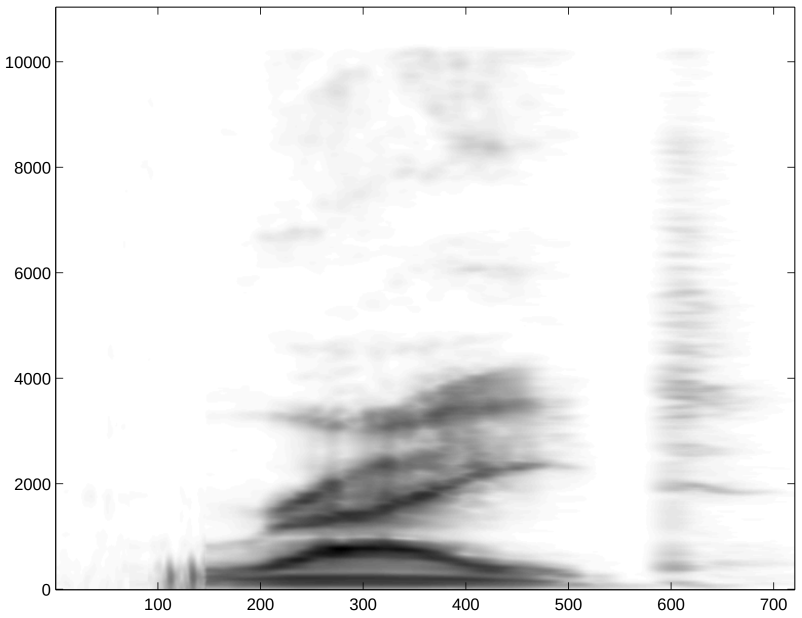


Figure 10: Spectrogram with a reduced 2nd-order structure.

3.5.2 Removing a second-order structure

The ‘remove 2nd’ button has been prepared to reduce the interference caused by the second-order structure. Even though the secondary-structure looks salient in a spectrographic display, it is difficult to detect reliably using frame-based methods. To avoid possible errors, it is common practice to embody *a-priori* knowledge in the processing stage. In the current implementation, a cepstrum lifter is introduced. The lifter is designed to attenuate cepstral components around $(2N+1)\tau_o/2$. In the term, N represents an integer, and τ_0 represents the fundamental period.

Figure 10 shows a processed spectrogram using the cepstral lifter. To get this image, please click the ‘removed spectrogram’ button. The F0 related horizontal structure is shown to be effectively removed.

The ‘bypass’ button simply skips this removal process.

3.5.3 Case 2: ‘analyze MBX’

Figure 10 also shows another deficiency in the previous STRAIGHT procedure. The spectrogram for plosive [t] around 600ms is temporally blurred by a relatively long time windowing. The long time windowing results from the extracted low F0 shown in Figure 5.

TEMPO usually tracks F0 even after the end of a voiced portion. It is not clear if there actually remains weak vocal fold vibration or if lower frequency noise (for example, air conditioning noise) has taken over.

‘analyze MBX’ provides a means to solve this problem. This part is rather new, meaning no fine tuning has yet been done. Please click the ‘analyze MBX’ button. Doing so will invoke a rather lengthy series of processes (listed below).

- Calculation of an adaptive spectrogram and calculation of a fixed window spectrogram which resembles the conventional wide-band spectrogram.
- Calculation of the complex auto-correlation at each 1/2 octave band.
- Calculation of the envelope auto-correlation at each 1/2 octave band.
- Calculation of the excitation allocation map in the time-frequency domain.

The final time-frequency representation is calculated based on the smoothed adaptive spectrogram (the second-order structure is removed using the ‘remove 2nd’ button) , the wide-band spectrogram, and the excitation allocation map.

Figure 11 shows the composite spectrogram as the final result. To get this image, please click the ‘final spectrogram’ button.

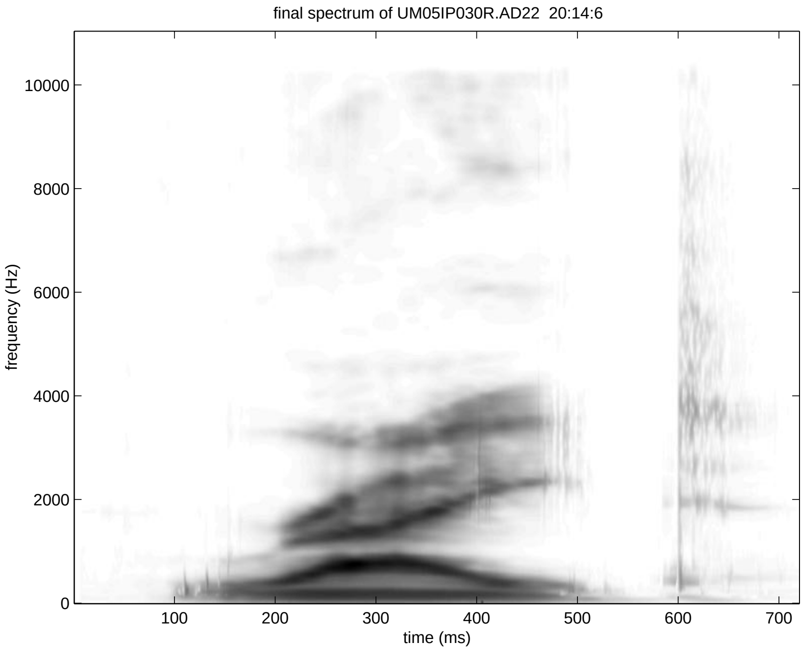
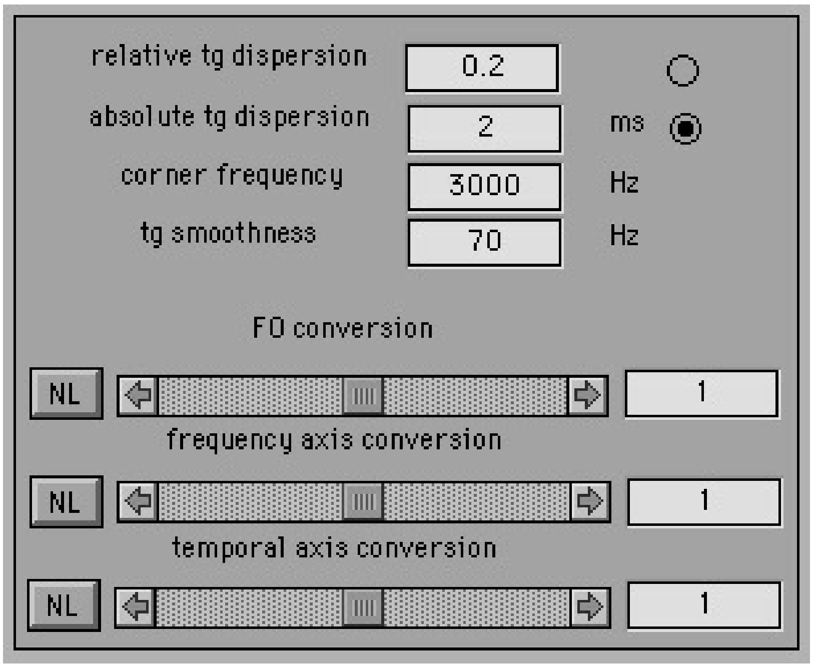


Figure 11: Composite spectrogram.



relative tg dispersion 0.2 ○

absolute tg dispersion 2 ms ●

corner frequency 3000 Hz

tg smoothness 70 Hz

F0 conversion

NL 1

frequency axis conversion

NL 1

temporal axis conversion

NL 1

Figure 12: Synthesis parameter control sub-panel.

3.6 Manipulation and re-synthesis (SPIKES)

The final task is to re-synthesize the manipulated[6] sound. The synthesis parameter control panel is illustrated in Figure 12.

[6] Note that manipulation in GUI-STRAIGHT is non-destructive. Manipulation in GUI-STRAIGHT is implemented as the modification of a mapping from an analyzed parameter to a synthesis parameter. This means that even after a large amount of manipulations, the base-line condition can be regained by resetting each mapping to the identity mapping.

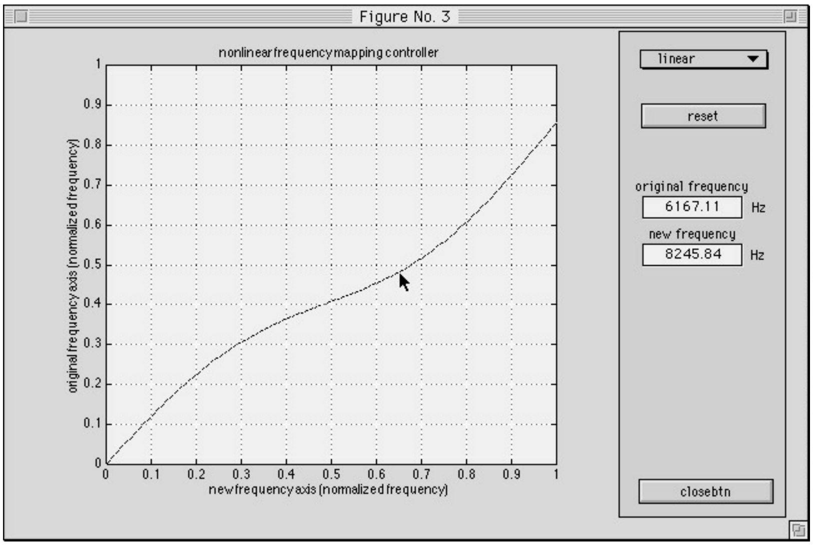


Figure No. 3
nonlinear frequency mapping controller
original frequency axis (normalized frequency)
new frequency axis (normalized frequency)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
reset
closebtn

linear

original frequency 6167.11 Hz

new frequency 8245.84 Hz

Figure 13: Non-linear frequency mapping controller.

3.6.1 Group delay design

A unique part of SPIKES is its group delay design to add artificial 'naturalness' to synthetic sound. The following parameters are open to manipulation by the user.

- **'relative tg dispersion' or 'absolute tg dispersion'** These alternatives determine the standard deviations of group delay dispersions. You can select the method for defining a dispersion. Based on our informal experience, using absolute values provides for more reasonable control than using relative values.
- **'corner frequency'** This parameter defines the boundary frequency between a stable region and a dispersion region. The transition between regions is a soft logistic function.
- **'tg smoothness'** This parameters provides a soft limit to the highest spatial frequency of a group delay shape. This approximately corresponds to the truncation of the Fourier transform of a group delay at $1/b_w$, where b_w is the number written in 'tg smoothness'.

Rewriting these parameters followed by 'CR' or 'ENTER' makes the system replace the old values with the new ones.

3.6.2 F0, frequency axis, and temporal axis manipulation

The lower part of the synthesis sub-panel provides for control over the pitch (F0 mapping), the vocal tract length (frequency axis mapping) and the speaking rate (temporal axis mapping). The controllers on the sub-panel are for proportional (linear) manipulations. Both the slider and edit box can be used to modify the same parameter.

The 'NL' buttons are for non-linear mapping. Clicking one of the 'NL' buttons gives you the corresponding non-linear mapping controller. Figure 13 shows an example of such a controller. Currently, only a frequency mapping interface is integrated. The interface design is very likely to be modified soon.

3.7 Re-synthesis

The 'synthsize grad' and 'synthesize' buttons are used to start the re-synthesis process. 'synthsize grad' invokes the synthesizer which uses a graded time-frequency map of the excitation source. 'synthesize' invokes the synthesizer with two-dimensional V/UV information.

3.8 Saving to file

The 'save to file' button is used to store the re-synthesized sound. After clicking this button, a file output graphical user interface pops out. The specific design of the user interface is system dependent.

Supported output file formats are the WAVE (.wav), old AIFF[7] (.aiff), and plain 16-bit linear binary format. Each file format is automatically determined based on the file name extension. No additional files are generated, but a description file output will be implemented in the near future.

[7] The old AIFF format is necessary, because the audio editing software system we are using does not recognize the AIFF-C format.

variable name	description
fs	sampling frequency in Hz
xold	input sound
f0raw	extracted F0 in Hz
nsgram	adaptive spectrogram
n2sgrambk	smoothed spectrogram
n2sgram	smoothed spectrogram with time domain enhancement
n3sgram	smoothed spectrogram (bypassed or 2nd-order removed)
f0var	error variance of F0
f0varL	error variance of F0 in a low frequency region
hbb	processed correlation map with τ_0 lag
sy	re-synthesized sound
delfrac	relative group delay dispersion as the standard deviation
delsp	absolute group delay dispersion as the standard deviation
cornf	corner frequency for group delay dispersion
gdbw	group delay smoothness in spatial frequency

Table 1: Variables and their descriptions.

4 For expert users

The ‘AUX’ sub-panel provides for direct control of internal variables. Clicking this button is equivalent to executing a ‘keyboard’ command inside an m-file function.

A list of variables is shown in Table 1.

To return to the normal operation mode, please type the following:

```
>>          return
```

You can also access these variables using ‘‘global’’ command of MATLAB in the command window, because these are global variables. The interface call-back procedure for GUI is helpful to find out how to call constituent procedures. The list of the procedure is in the appendix.

5 Request for your comments

I would definitely appreciate your comments and feedback. Please send mail to

```
kawahara@sys.wakayama-u.ac.jp
```

if you have anything to share. If it is possible for you to send files that reveal deficiencies in STRAIGHT, this would be is very helpful in improving future versions.

A Call-back program list

See the following pages.