GUI-STRAIGHT: Getting started

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

draft 10:21 P.M., February 26, 1998

Contents

1 Introduction

STRAIGHT-suite[1] is a set of procedures to analyze, modify and synthesis speech-like sounds. Recent introduction of the GUI-STRAIGHT[2] made it extremely easy to use STRAIGHT. This document provides a step-by-step introduction of this GUI-STRAIGHT.

2 System requirements

Current version of GUI-STRAIGHT is reported to operate on the following platforms:

2.1 Software requirement

MATLAB version 5.0 or later and signal processing toolbox is recommended. No other toolboxes are not necessary to run GUI-STRAIGHT. Some component procedures of STRAIGHT may be functional on older version of MATLAB, however it is strongly recommended to use version 5.0 or later.

For platforms other than Macintosh, it is also recommended to have an audio input function which is called inside MATLAB. It is still possible to use a separate software 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 and Power Book.**
- **[Windows95] PC with Pentium and Pentium II.**
- **[UNIX] Sparc, SGI and HP.**
- **[linux]**

2.2.1 Machine power

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

2.2.2 Memory requirements

Larger the better. :-)
Sometimes, 64MB is not enough.

3 Tour

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

3.1 Search path

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

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

In this example, STRAIGHT programs are located in the 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 references listed in the end of this document.

[2] GUI stands for Graphical User Interface, as you expect.

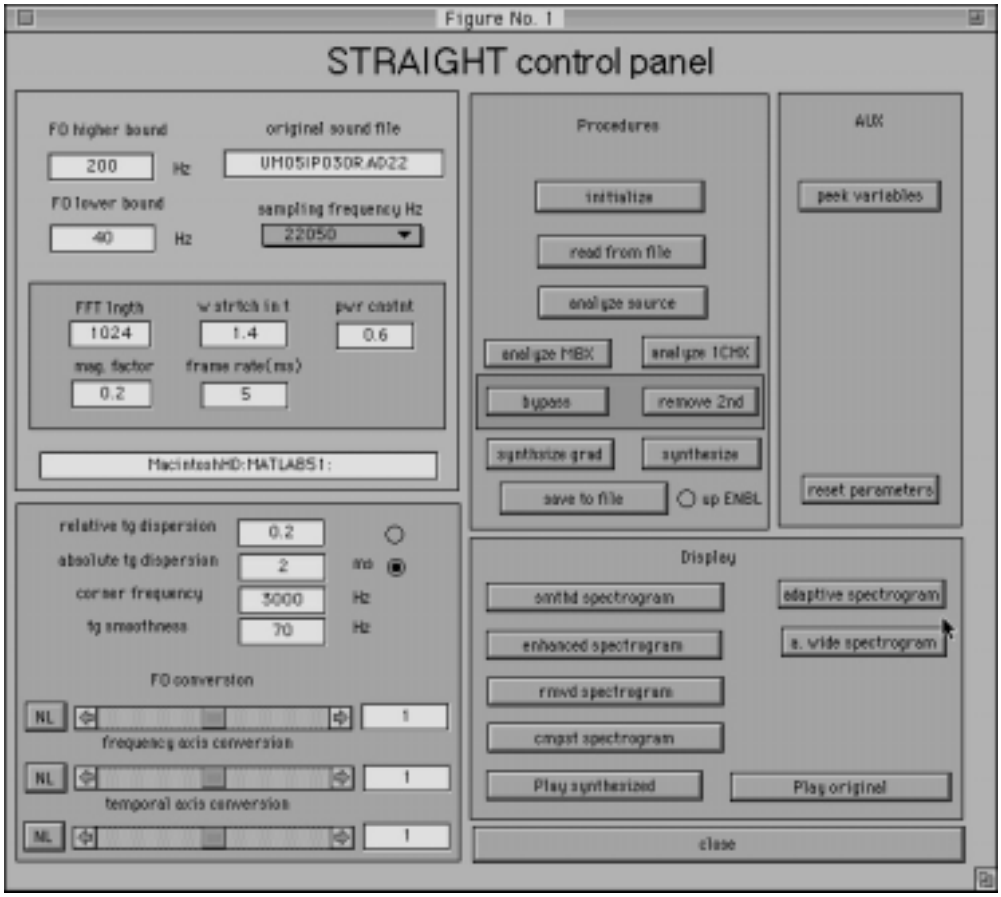

Figure 1: STRAIGHT control panel.

3.2 STRAIGHT control panel

Now, please type the command

```
>> straight
```

to start the system.

Then the control panel (Figure. 1) will be shown in a couple of seconds. Buttons which are meaningful at each context are made active. Inappropriate buttons are disabled. The first button to be clicked is 'read from file' button.

The following subsubsections briefly introduce which is which.

3.2.1 Procedure panel

Upper center panel is for general procedures. 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 courses for processing sounds. As shown in the figure, two lines of procedures are supported at present. The course via 'analyze 1CHX' uses a couple of single vectors to carry V/UV (voiced/unvoiced) information. The other course via 'analyze MBX' uses a V/UV map to represent periodicity in each time-frequency region.

There are two alternatives for each course. One is to use the analysis result directly in manipulation and re-synthesis. The other removes the spectral second-order structure[3] before manipulation and re-synthesis.

3.2.2 Display panel

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

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

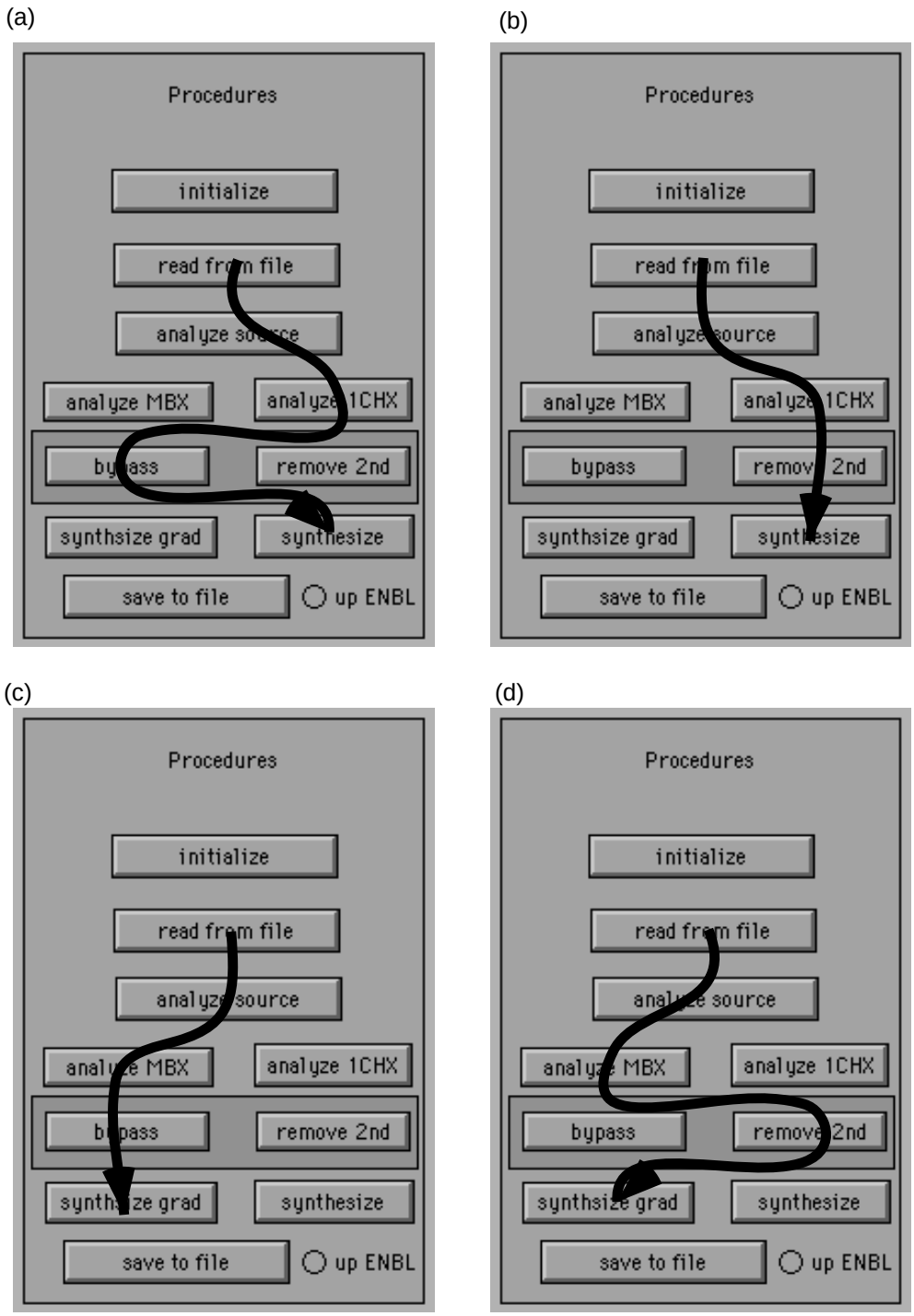

Figure 2: Alternative courses for manipulating sounds.

3.2.3 AUX panel

Miscellaneous functions will be placed on this panel.

3.2.4 Analysis parameter panel

Parameters mainly used in the analysis stage are made accessible and controllable using MATLAB 'edit' and 'menu selection' primitives.

3.2.5 Manipulation and synthesis panel

Parameters used in the synthesis stage are made accessible and controllable using MATLAB 'edit', 'slider' and 'radio button' primitives. Nonlinear arbitrary mapping of the original and the synthesis parameters

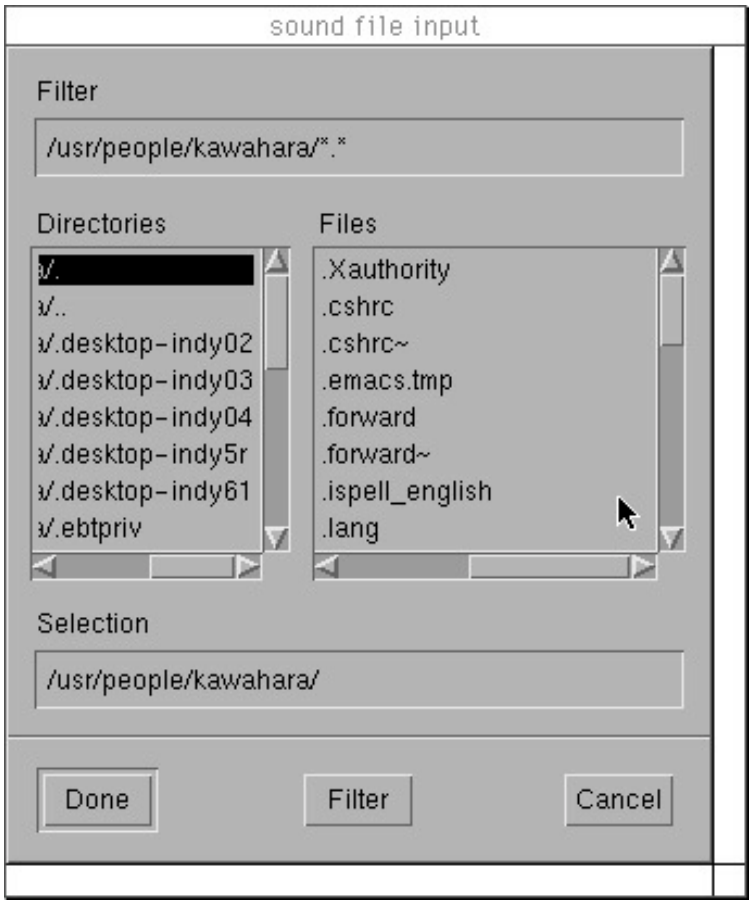

Figure 3: File input user interface for UNIX.

can be controlled using a direct manipulation controller.

3.3 Reading speech file

The next step is to read a speech file to manipulate. By clicking 'read from file' button, you will get the user interface for file input. The usual dialogue file interface pops up for Macintosh environment. Figure 3 shows the graphical file interface for UNIX environment.

The file input routine can understand the WAVE format, the AIFF (and AIFF-C) format and the plain 16bit linear binary. WAVE and AIFF processing are 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 unknown extension are assumed as the plain binary.

After reading files with header information, sampling frequency will be automatically replaced by the value read from the file. For plain binary files, users have to set the sampling frequency manually.

3.4 Analyzing source information (TEMPO)

Since STRAIGHT is a pitch adaptive procedure, it is crucially important to extract F0 information reliably for the first time. By clicking 'analyze source' button, source information extraction using TEMPO is invoked. But, please read the following subsection before clicking the button.

3.4.1 Notes on analysis parameters

GUI-STRAIGHT provides a mean to tweak analysis parameters using 'analysis parameter control sub-panel', which is shown in Figure 4. To speed up the whole process, some parameters are better to be modified from the original value. 'F0 lower bound' and 'F0 higher bound' define the F0 search range in pitch extraction using TEMPO. If there are *a-priori* knowledge about possible F0 is available, using the information to trim this search range eliminates unnecessary processing.

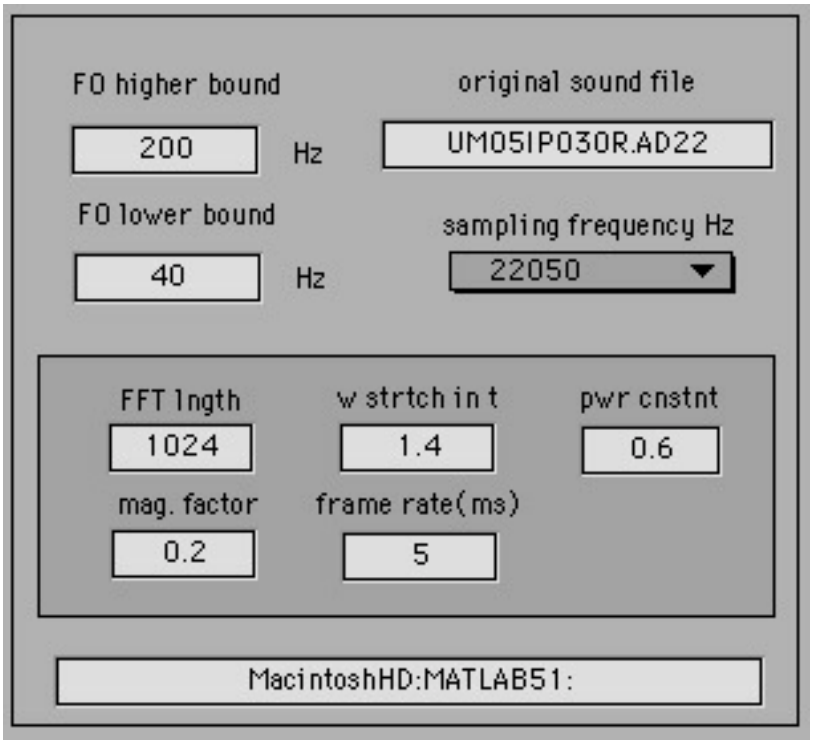

Figure 4: Analysis parameter control sub-panel. (Upper left of the STRAIGHT control panel.)

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

Sampling frequency control is implemented as a 'popup menu'. If the file under investigation is either 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 in Hz.

Other parameters shown on the sub-panel have following functions.

- **FFT length** This parameter is automatically set, based on the sampling frequency. Internal FFT uses this length as the frame length. The length is the smallest 2^N (where N is an integer.) to cover 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 window.
- **pwr cnstnt** This should be spelled out as "power constant". This determines a nonlinear mapping function $g(x)$ for smoothing operation. Specifically, the nonlinear function of absolute spectral value x has the following form.

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

  where α represents power constant. The default value $\alpha = 0.6$ approximate loudness-preserving smoothing operation.
- **mag. factor** This represents magnification factor in the time domain. This factor was introduced in my ASJ technical meeting article in July 1997. This time domain operation enhances sharpness of the spectral peaks. However, in a retrospective consideration, this parameter is excessive at least for the current implementation of STRAIGHT-core. Increasing this value adds some 'richness in resonance', but also may introduce an artificial timbre. The default value 0.2 is tentative. Setting 0 for this factor makes 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 eliminated the need of temporal smoothing.

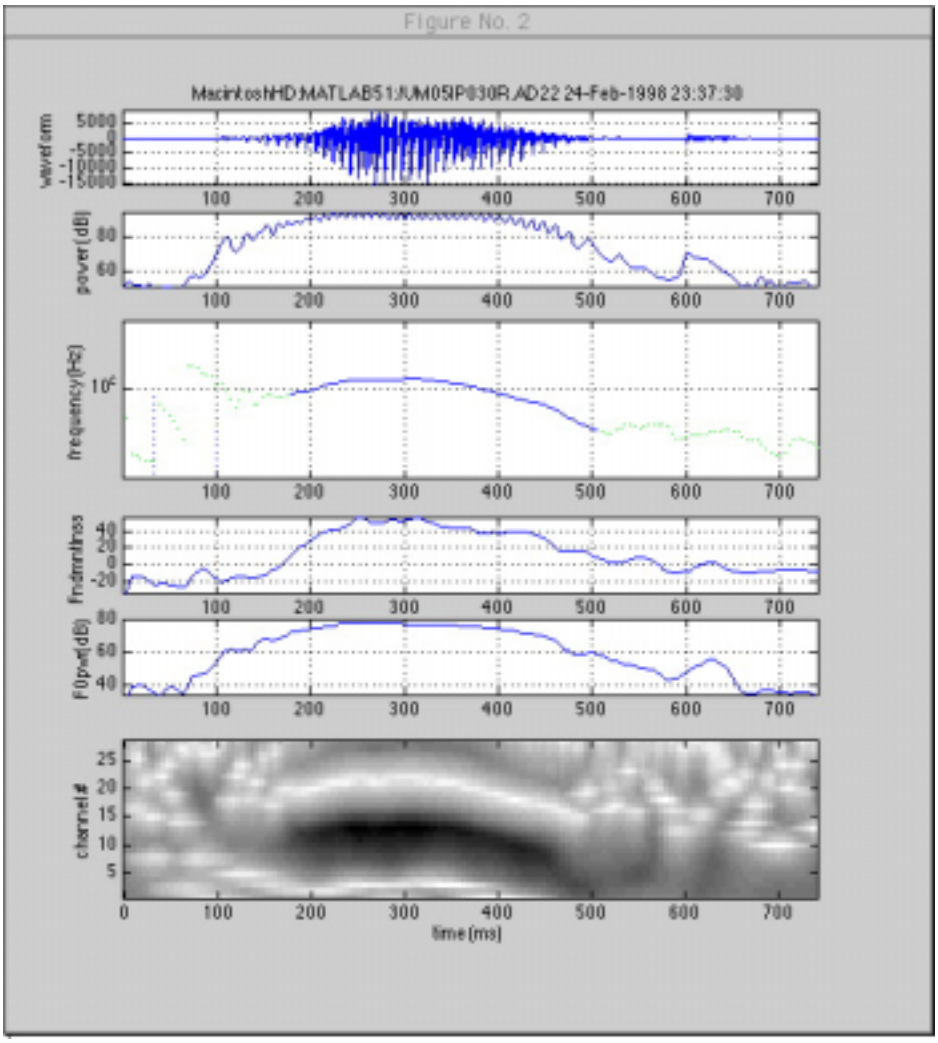

Figure 5: Extracted source information for a male utterance 'right'. (from top to bottom: waveform, power in the F0 search rage, extracted F0, 'fundamentalness', F0 power and 'fundamentalness' map.)

3.4.2 Source information display

After clicking the 'analyze source' button, it will take several minutes to calculate necessary source information for later STRAIGHT processing.

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

The fourth plot shows the highest 'fundamentalness' at each frame. This value is roughly 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 the channel-1 corresponds[5] to 'F0 lower bound' defined in the analysis sub-panel. The channel interval is set to $2^{1/12}$.

3.5 Extracting spectral envelope (STRAIGHT-core)

Next step is spectral analysis. Please click 'analyze 1CHX' button or 'analyze MBX' button to start the STRAIGHT-core procedure. After completion of the procedure, 'bypass' and 'remove 2nd' button are made

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

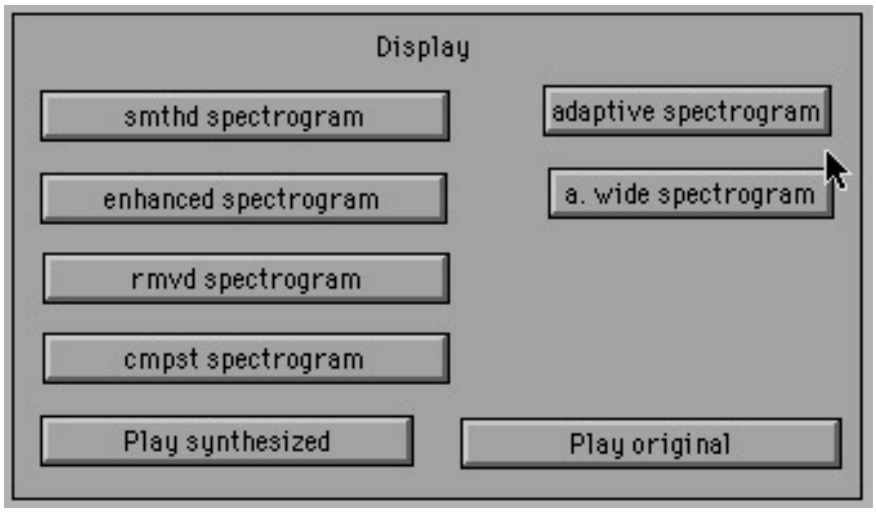

Figure 6: Display control sub-panel.

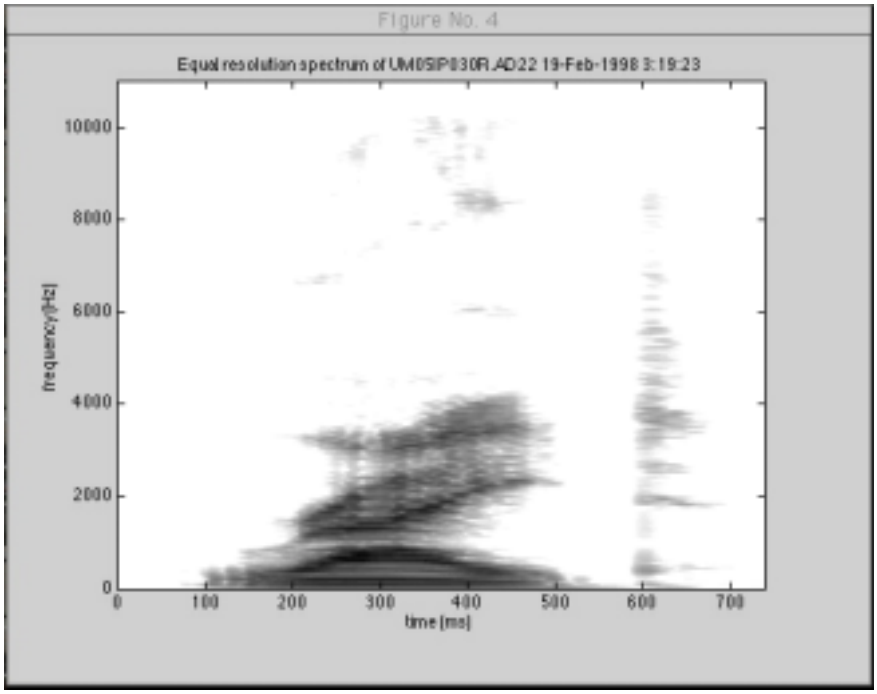

Figure 7: Spectrogram with a compensatory set of adaptive time windows.

active. Other appropriate buttons on display sub-panel are also activated.

3.5.1 Case 1: 'analyze 1CHX'

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

Figure 7 shows an adaptive spectrogram based on the F0 information of TEMPO output. To get this image, please click 'adaptive spectrogram' button. Since resolution in the frequency domain is set higher than that in the time domain, the harmonic structure is resolved. It is also noted that even without the temporal smoothing in the original STRAIGHT, spectral variations in the time domain is 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 1980s in his speech coding papers. This regular spectral structure implies that there is an additional excitation at precise center position be-

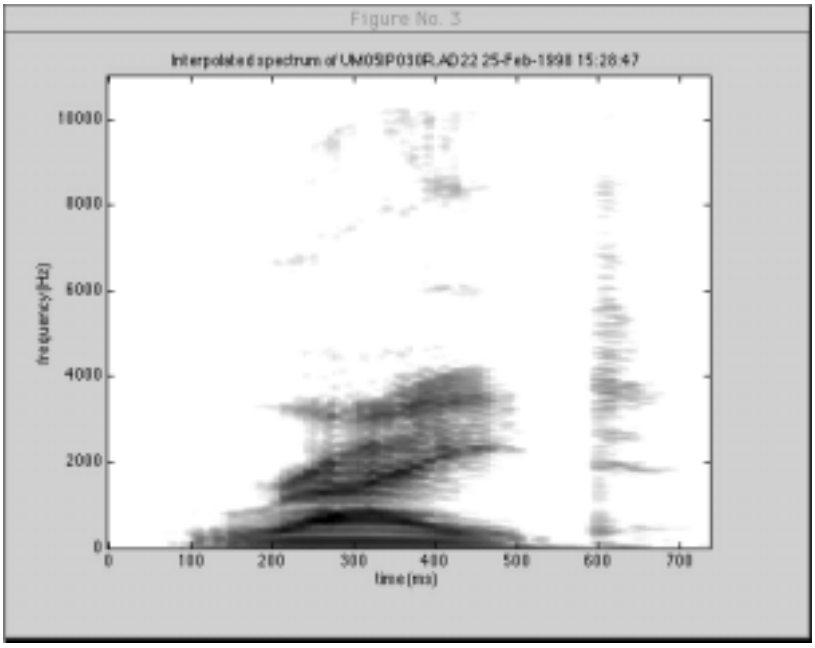

Figure 8: Smoothed and optimally recovered spectrogram. STRAIGHT-core smoothing in the time domain with an optimal smoothing function is employed.

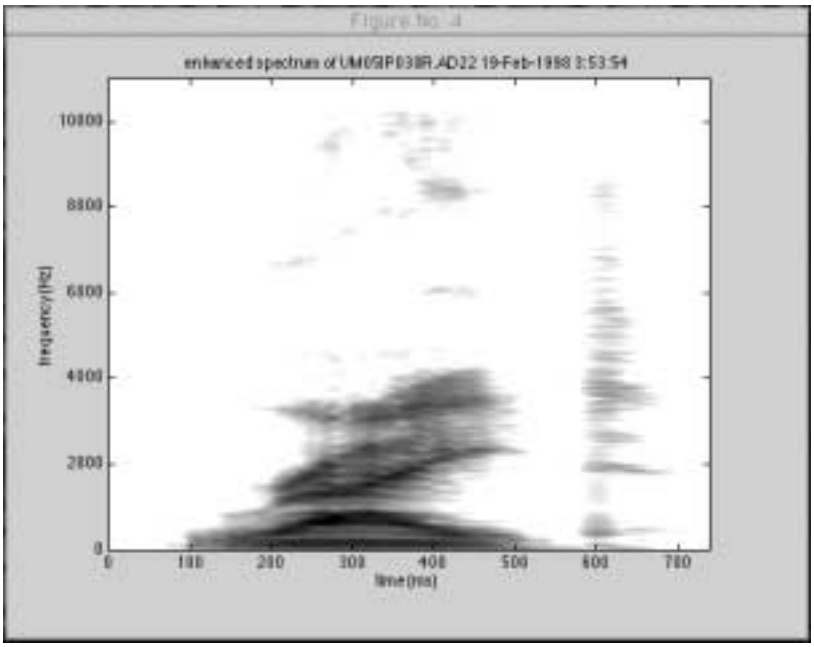

Figure 9: Enhanced spectrogram made from the smoothed spectrogram.

tween adjacent primary excitation pulses. 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 'smthd spectrogram' button.

Figure 9 shows an enhanced spectrogram using a time domain processing. To get this image, please click 'enhanced spectrogram' button.

Both spectrograms illustrates 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 also found in a female utterance.

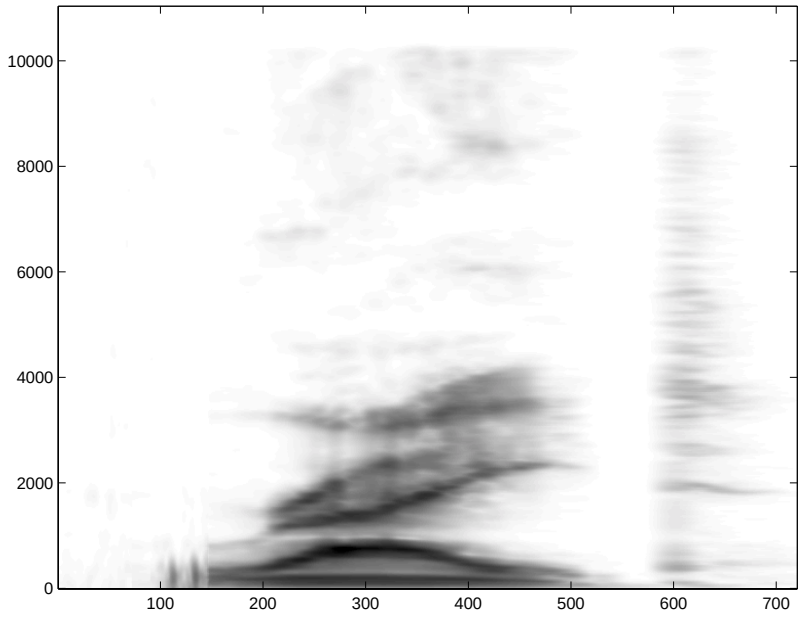

Figure 10: Spectrogram with reduced 2nd-order structure.

3.5.2 Removing second-order structure

'remove 2nd' button is prepared to reduce the interference by the second-order structure. Even though the secondary-structure looks salient in a spectrographic display, it is difficult to detect the structure reliably using frame-based methods. To avoid possible errors, it is a 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 $(2N+1)\tau_o/2$. Where N represents integers, τ_0 represents fundamental period.

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

'bypass' button simply skips this removal process.

3.5.3 Case 2: 'analyze MBX'

Figure 10 also shows the other deficiency in the previous STRAIGHT. Spectrogram for plosive [t] around 600ms is temporally blurred by a relatively long time windowing. The long time windowing is resulted from 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 lower frequency noise (for example air conditioning noise) had taken over.

'analyze MBX' provides a mean to solve this problem. This part is rather new, meaning no fine tuning was done yet. Please click 'analyze MBX'. It will invoke a rather lengthy series of processes listed below.

- Calculation of an adaptive spectrogram and calculation of fixed window spectrogram which resembles the conventional wide-band spectrogram.
- Calculation of complex autocorrelation at each 1/2 octave band.
- Calculation of envelope autocorrelation at each 1/2 octave band.
- Calculation of 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 '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 'final spectrogram' button.

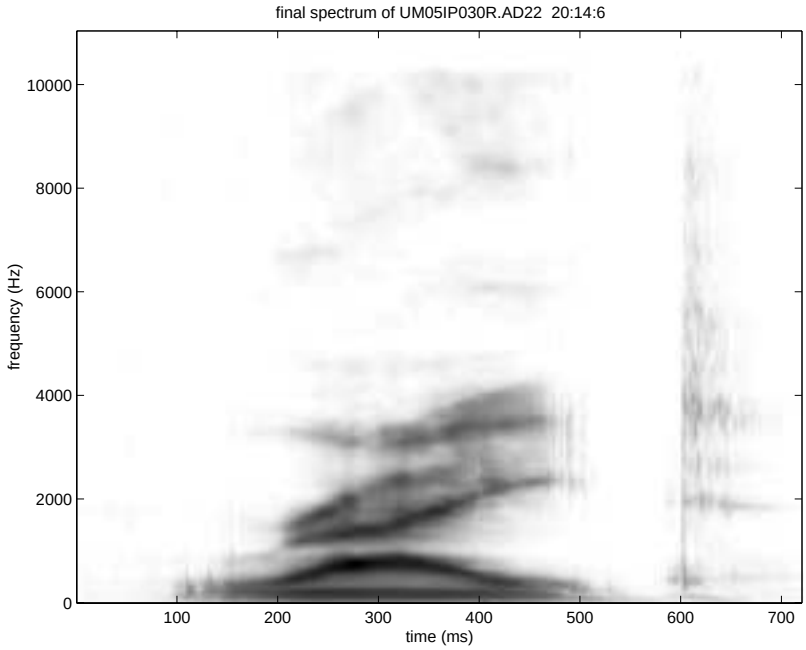

Figure 11: Composite spectrogram.

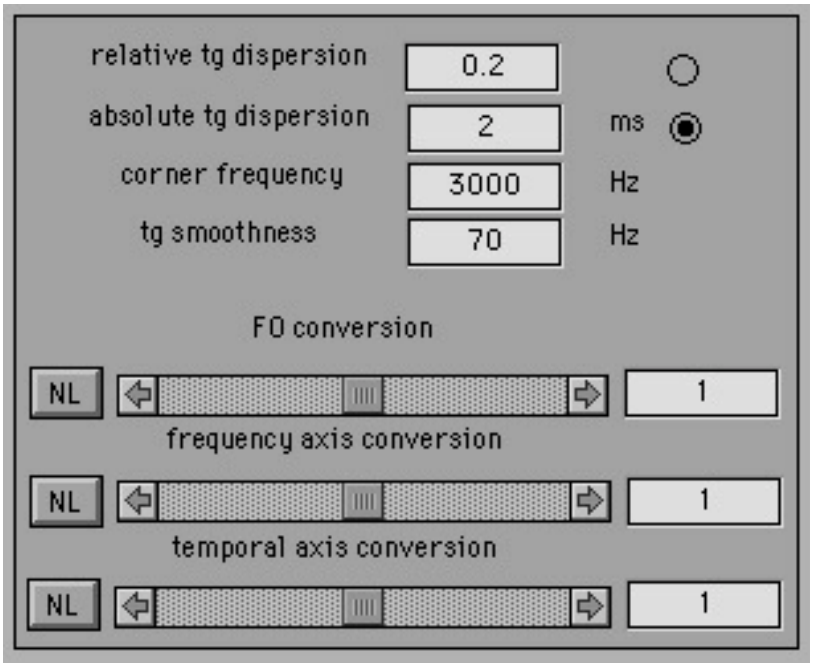

Figure 12: Composite spectrogram.

3.6 Manipulation and re-synthesis (SPIKES)

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

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

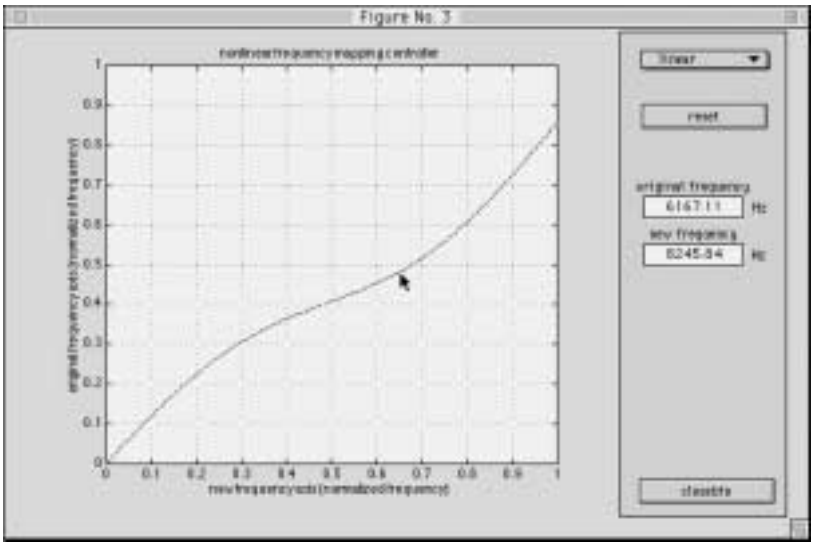

Figure 13: Non-linear frequency mapping controller.

3.6.1 Group delay design

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

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

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

3.6.2 F0, frequency axis and temporal axis manipulation

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

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

3.7 Re-synthesis

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

3.8 Saving to file

'save to file' button is used to store the re-synthedsized sound. 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 WAVE (.wav), old AIFF[7] (.aiff) and plain 16 bit-linear-binary. 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]Old AIFF format was necessary, because the audio editing software system we are using does not recognize 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 the time domain enhancement
n3sgram	smoothed spectrogram (bypassed or 2nd-order removed)
f0var	error variance of F0
f0varL	error variance of F0 in low frequency region
hbb	processed correlation map with τ_0 lag
sy	re-synthesized sound
delfrac	relative group delay dispersion as standard deviation
delsp	absolute group delay dispersion as standard deviation
cornf	corner frequency for group delay dispersion
gdbw	group delay smoothness in spatial frequency

Table 1: Variables and their contents.

4 For expert users

'AUX' sub-panel provides direct control of internal variables. Clicking this button is equivallent to execure 'keyboard' command inside a m-file function.

A list of variables is shown in Table 1.

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

```
>> return
```

5 Request for your comments

I appreciates 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 the file which revealed deficiensy of STRAIGHT, it is very helpful.