Faster than real-time and audio sampling rate extraction of fo and event candidates using an analytic signal with a cosine series envelope

(compiled: 5:33am December 25, 2018)

Hideki Kawahara[1], *Ken-Ichi Sakakibara*[2], *Masanori Morise*[3]

[1]Wakayama University, Wakayama Japan
[2]Health Science University of Hokkaido, Sapporo Japan
[3]University of Yamanashi, Kofu Japan

kawahara@sys.wakayama-u.ac.jp, quesokis@gmail.com, mmorise@yamanashi.ac.jp

Abstract

We introduce a set of algorithms and tools for representing local periodicity and event structure using an analytic signal with a specialized cosine series. The primary use of them is for designing the excitation source signals of the next generation VOCODERs. The procedure uses efficient instruction sets of vector math libraries, which enabled replacing Flanagan's equation with more direct implementation of the instantaneous frequency and the group delay. The algorithms run faster than real-time and yields fo candidates and event candidates at the audio sampling rate resolution.

One implementation of periodic structure analysis uses an analytic signal with close to isotropic time-frequency resolution for wavelet-based filter arrangement. Integrating the squared group delay and squared temporal derivative of the group delay provides the measure for representing the salience of the periodicity and is proportional to signal to noise ratio and consequently the variance of the frequency.

The other implementation of event structure analysis uses the similar wavelet arrangement of an analytic signal. Combining with an inverse filter for removing responses corresponding resonant frequencies of vocal tract enables clear representation of discontinuities corresponding to excitation events.

We also made tools for real-time visualization of these attributes using MATLAB implementation and off-line tools for detailed analyses. These tools and supporting functions are available as open-source software on GitHub.

Index Terms: speech analysis, speech synthesis, excitation source, periodicity, events, instantaneous frequency, group delay

1. Introduction

We propose an extension of speech signal representation by introducing a generalized tool for periodicity and event analyses. This extension enables analysis and synthesis of less explored aspects of speech signals, excitation source. These excitation attributes enrich speech especially adding para- and non-linguistic contents to speech sounds. These algorithms and tools are designed for fundamental investigation of excitation source attributes. We ask the following fundamental questions. a) Is fo always relevant for representing excitation regularity? b) Is there any perceptual structure for representing irregular vocal fold vibration? c) What is the most relevant description of subharmonic behavior of excitation? d) How to construct excitation signals for reproducing these non-typical excitation sources to stimulate similar perception? e) How to contribute discussion [1] on the terminology of voice attributes?

2. Background

Instantaneous frequency and group delay are useful for analyzing the excitation source of voiced sounds. We introduce a set of simple implementations for calculating these values. Historically their implementations used Flanagan's equation [2] and its variant because those equations do not require phase unwrapping nor inverse trigonometric functions, which have been fragile and inefficient operations.

Recent advances of CPUs and GPUs changed this situation. They provide specialized instruction sets and libraries for those operations (for example, refer to [3]). This situation makes simple, more direct implementations based on the definitions of instantaneous frequency and group delay practical substitutes of the Flanagan's equation.

Phase information is more sensitive to discontinuities (in value, derivative, and higher order derivatives) than power information. We found that a cosine series introduced for deriving closed-form representation of antialiased L-F and F-L models [4, 5] is also relevant for analyzing signal phase and derived attributes [6]. Combining this cosine series and the previously mentioned simple implementation resulted in a procedure that is faster than real-time for calculating those attributes.

3. Instantaneous frequency and group delay

Instantaneous frequency $\omega_i(t)$ of a complex valued signal $x(t) = |x(t)| \exp(j\theta(t))$ is the time derivative of its phase function $\theta(t)$, where $j = \sqrt{-1}$.

$$\omega_i(t) = \frac{d\theta(t)}{dt} \tag{1}$$

Group delay $\tau_g(\omega)$ of a complex valued function $x(\omega) = |x(\omega)| \exp(j\theta(\omega))$ is the (angular) frequency derivative of its phase function $\theta(\omega)$ with the negative sign.

$$\tau_g(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{2}$$

3.1. Flanagan's equation

By taking logarithm of the $x(t)$ and applying derivative rules yields the following equation to calculate the instantaneous frequency.

$$\omega_i(t) = \frac{\Re[x(t)]\Im\left[\frac{dx(t)}{dt}\right] - \Re\left[\frac{dx(t)}{dt}\right]\Im[x(t)]}{|x(t)|^2} \tag{3}$$

where $\Re[x]$ and $\Im[x]$ represent the real and the imaginary part of x, respectively. Equation 3 is the Flanagan's equation

[2]. Substituting ω with t and adding the negative sign yields the similar equation for the group delay. These equations do not require calculations of phase unwrapping nor inverse trigonometric functions.

3.2. Simple discrete implementation

Argument of the ratio of succeeding samples of a discrete signal $x[n]$ is proportional to the instantaneous frequency $\omega_i[n]$.

$$\omega_i[n] = \angle \left[\frac{x[n+1]}{x[n]} f_s \right], \tag{4}$$

where f_s represents the sampling frequency.

Similarly, argument of the ratio of neighboring samples of a discrete spectrum $X[k]$ is proportional to the group delay $\tau_g[k]$.

$$\tau_g[k] = -\frac{1}{\Delta\omega} \angle \left[\frac{X[k+1]}{X[k]} \right], \tag{5}$$

where $\Delta\omega$ represents the spacing of the discrete (angular) frequencies $\omega[k+1]$ and $\omega[k]$.

3.3. Analytic signal

We used a cosine series envelope of the analytic signal for analyzing the phase of the signal. We found it is necessary to use a relevant envelope when analyzing phase of the signal [6]. The following equation defines the envelope and the analytic signal using the envelope.

$$w_e[n; c_{\mathrm{mag}}, N] = \sum_{m=0}^{M} a_m \cos\left(\frac{2\pi n m}{c_{\mathrm{mag}} N}\right) \tag{6}$$

$$w[n; c_{\mathrm{mag}}, N] = w_e[n; c_{\mathrm{mag}}, N] \exp\left(\frac{2j\pi n}{N}\right) \tag{7}$$

where $M = 5$ represents the highest order of the cosine series, and the support of this function is $[-c_{\mathrm{mag}}N, c_{\mathrm{mag}}N]$. The coefficients of the six-term series are 0.2624710164, 0.4265335164, 0.2250165621, 0.0726831633, 0.0125124215, and 0.0007833203 from a_0 to a_5. The sidelobes have the highest level of -114 dB and the decay rate of -54 dB/oct.

4. Cost functions

We revisit the physical meaning of instantaneous frequency and group delay [7]. Revisiting them provides simple and computationally efficient candidates of cost functions. In this section, we propose to use these cost functions for evaluating candidates of salient frequency and salient event derived from fixed-point analyses [4, 8]. (Actually, the referred articles essentially used these cost functions, in retrospective views.)

5. MATLAB functions

We implemented these functions using MATLAB. We also prepared test scripts for the calibration and evaluation of these implementations. The core functions are as follows:

designCos6Wavelet.m Function to design a set of analytic signals for wavelet analysis.

waveletSourceAnalyzer.m Calculates wavelet analysis results using FFT-based efficient convolution.

sourceInformationAnalysis.m Calculates signal attributes using `waveletSourceAnalyzer.m`.

staticTrigBsplinePowerSpec.m Calculates interference-free power spectral representation using fo information [9].

This release provides test scripts for these functions in "test" directory. Appendix A shows excerpts of the test results using scripts. The `sourceInformationAnalysis.m` runs faster than real-time even when calculating fo candidates at the audio sampling rate, for example 44,100 Hz. It runs 12 times faster than real-time with automatic downsampling.

6. Tools

We introduced a set of tools for real-time interaction with wavelet analyses and post-processing tools. This section provides their roles and a brief introduction of each tool.

6.1. Real-time visualization of wavelet analysis

These tools are for observing the behavior of signal attributes using wavelet analysis. It does not provide any decision results other than the fundamental component. The fundamental component extraction uses an ad hoc stability measure and provides stabilization of waveform display and fine-tuning feedback.

Figure 1 shows the GUI of the real-time visualizer of wavelet analysis. It visualizes the output of a filter bank consisting of the analytic signal impulse response. The GUI has four panels. From top left with counterclockwise, the panels show a) input waveform, b) phase map of the filter output, c) power of filter outputs, and d) fo fine tuner with closest musical note name. It also has several UI controllers. They are "START," "STOP," "SAVE," and "QUIT" buttons, and popup menus for selecting display information (phase, amplitude, instantaneous-frequency, and group delay) and operation modes (Normal or Experimental). The bottom image of Fig. 1 shows the "Phase" information in the "Experimental" mode.

6.1.1. Waveform

The top left panel shows the waveform which corresponds to the center location of the map display below. The waveform center corresponds to the assigned phase of the fundamental component of the signal. The edit box UI controller provides a way to define the assigned phase. The default assigned value is -120 degree, which roughly corresponds to the GCI (Glottal Closure Instant).

6.1.2. Attribute map

The lower left panel of each image shows the time-scale map of attributes. Currently, four types of map are available; phase map, amplitude map, instantaneous frequency (deviation) map, and group delay map. This panel has a chromatic scale axis at the right side having green lines indicating the position with a semitone step. It also has musical note names.

The top image of Fig. 1 shows the phase map. The phase map uses "hsv" colormap of MATLAB because the phase and the hue of color share the same cyclic topology.

The middle image of Fig. 1 shows the amplitude map. Since our auditory system has a wide dynamic range in sound level, the pseudo color mapping uses dB representation of the amplitude. This value has a monotonic topology and uses "jet" colormap of MATLAB.

The top image of Fig. 2 shows the normalized instantaneous frequency map. The corresponding center frequency normalized each instantaneous frequency. The topology of this attribute is linear, but the value has a chance to be singular. This map uses "jet" colormap of MATLAB. This map uses logarithmic conversion for mapping to pseudocolor and used truncation inside 1/2 to 2. The color green corresponds to zero.

The bottom image of Fig. 2 shows the normalized group delay. The group delay of each filter uses its center frequency

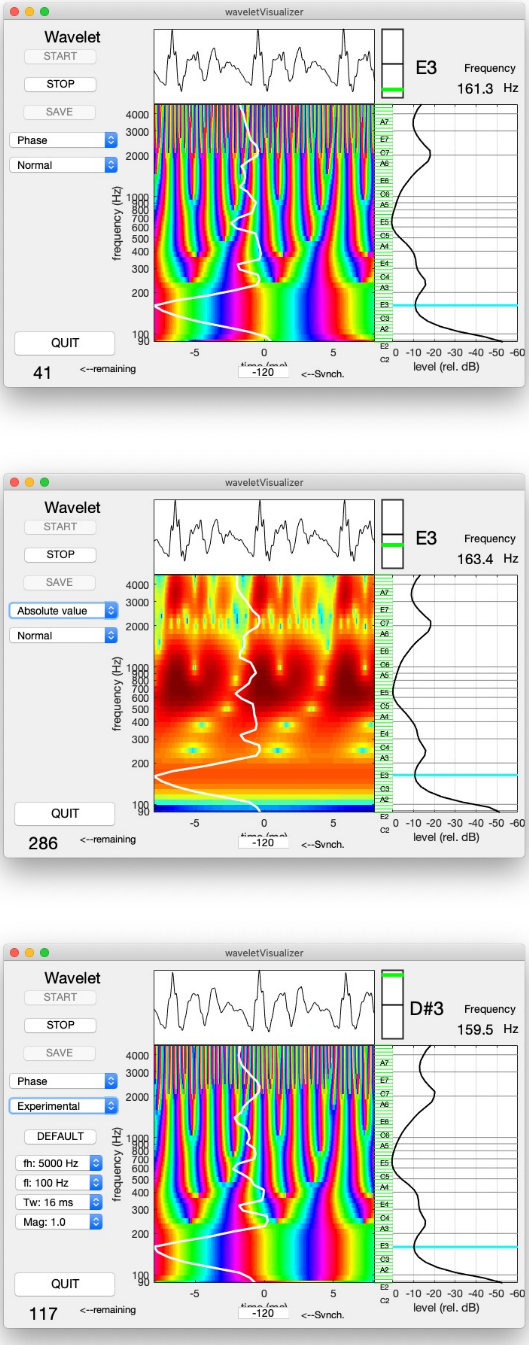


Figure 1: *Realtime wavelet visualizer. From top to bottom, phase view, dB absolute value view and phase view with extra control parameters.*

to normalize the value. The topology of this attribute is linear, but the value has a chance to be singular. This map uses "jet" colormap of MATLAB. Since the time axis is linear, we used no logarithmic conversion but used truncation from -0.75 to 0.75. The color green also corresponds to zero.

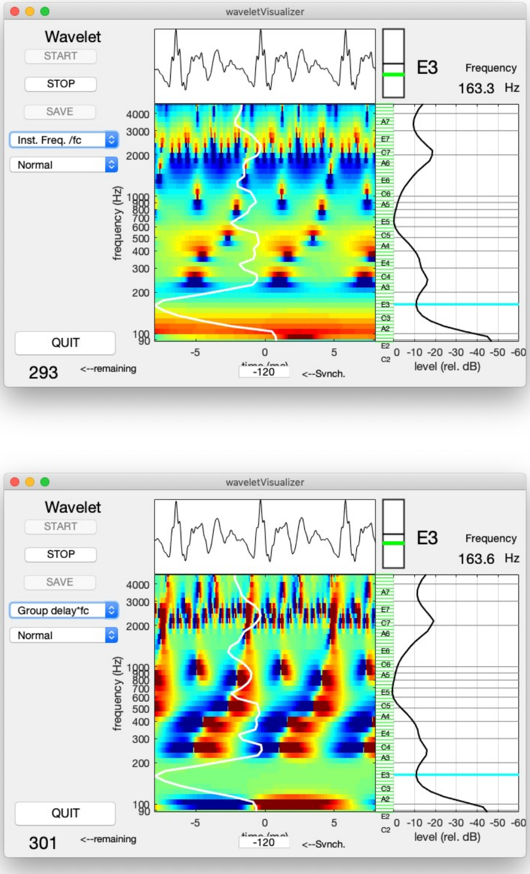


Figure 2: *Realtime wavelet visualiser. Upper image shows the normalized instantaneous frequency map and the lower image shows the normalized group delay map.*

6.1.3. *Level and fo indicator*

The right bottom plot of each image shows the averaged power of each filter output. The duration of averaging uses the segment shown in the waveform display and the attribute map display. The cyan line indicates the frequency of the most salient periodic component. Usually, in a voiced segment, it corresponds to conventional fo.

6.1.4. *Tuning indicator*

The top right corner shows the frequency of the most salient periodic component numerically and visually. The left indicator shows the corresponding musical note name and the deviation of the salient frequency from the reference frequency of the musical note (in equal temperament with the 440 Hz reference). The vertical span of the box is 100 musical cent, and the horizontal black line indicates the frequency of the musical note. The green horizontal line shows the average frequency of the most salient periodic component.

6.1.5. *Experimental mode*

The left popup menu provides to change "Normal" mode to "Experimental" mode. In experimental mode, the user can change the following analysis settings: a) High-end frequency

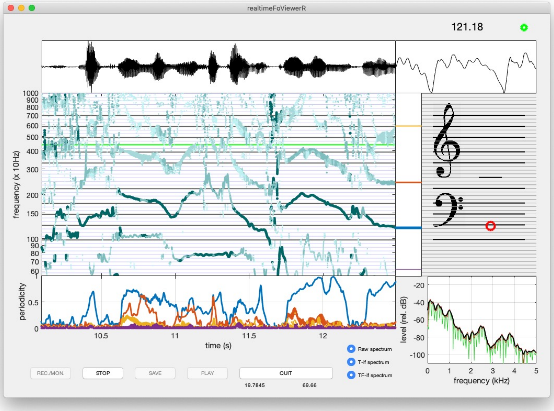
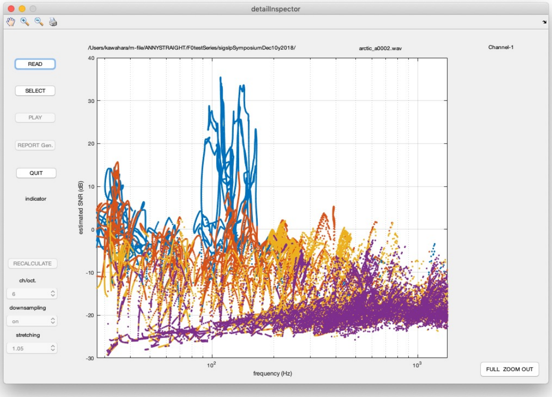


Figure 3: *Realtime visualization of fo candidates.*



/Users/kawahara/m-file/ANNYSTRAIGHT/F0testSeries/sigslpSymposiumDec10y2018/

arctic_a0002.wav

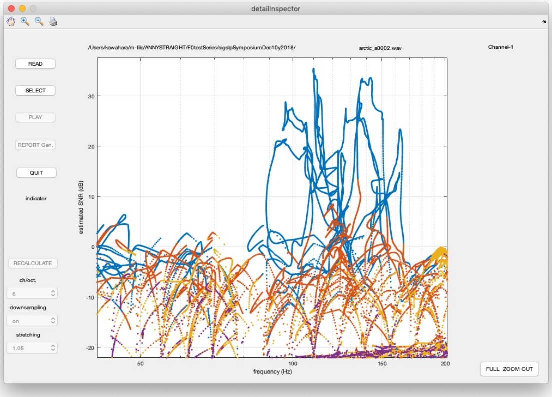
Channel-1

ch/oct. 6

downsampling on

stretching 1.05

Figure 4: *Scatter distribution of periodicity candidates on the frequency vs. periodicity measure plane.*



/Users/kawahara/m-file/ANNYSTRAIGHT/F0testSeries/sigslpSymposiumDec10y2018/

arctic_a0002.wav

Channel-1

ch/oct. 6

downsampling on

stretching 1.05

Figure 5: *Scatter distribution of periodicity candidates on the frequency vs. periodicity measure plane. This shows just after region selection.*

of periodicity check, b) Low-end frequency of periodicity check, c) Signal and map viewing width, and d) Temporal stretching factor of the envelope (1 indicate isotropic resolution both in the time and the frequency domain.)

6.2. Real-time fo candidate analysis

Figure 3 shows the GUI of the real-time extractor of fo candidates. The left three panels are horizontally scrolling viewers of (from top to bottom) waveform, fo candidates, and periodicity indicator. They are updated every 50 ms. The top right plot shows the stabilized view of the waveform. The center of the view locks to the zero-phase of the most salient periodic component. Usually, it is the fundamental component. The middle right panel shows the frequency of the most salient component on the musical score-like representation using a red circle. The G and F clefs represent the reference points. The panel has gray horizontal lines representing the chromatic scale (semitone step). The bottom right panel shows the interference-free spectral representations [9] and usual power spectrum using the Blackman window.

The tool keeps running endlessly. The scrolling buffer length is 30 s. When reaching to the end, it initializes and starts again. The user can stop running by clicking the “STOP” button. Then, by using the “PLAY” button, the user can replay the contents of the buffer. By clicking the “SAVE” button, the user can save the contents in the current buffer. The tool generates a unique name for recording the contents. The “START” button restarts the tool.

We designed this tool also for voice therapy and training. The periodicity panel in the bottom provides qualitative feedback about the salience of the voice periodicity. It uses a heuristic conversion to map periodicity into 0 to 1 range (0: completely random, and 1: purely periodic). We are trying to find the better heuristic conversion for this panel.

6.3. Post-processing tool for detailed analysis

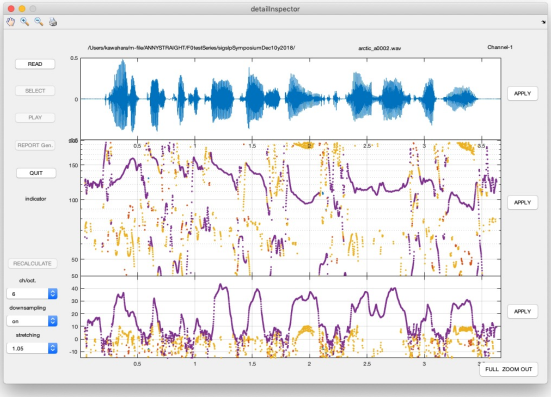
We introduced an interactive tool for analyzing speech signals using the analysis functions of signal attributes. The faster than real-time analysis procedures enabled flexible interaction.

Figure 4 shows the scatter plot of the extracted periodicity candidates on the frequency vs. periodicity measure (converted to SNR) plane. The tool displays this view just after reading a speech file. The user can select the region where the relevant fo likely exists by using a MATLAB UI tool on the top left corner of the GUI. The tested signal is an excerpt from the CMU ARCTIC database [10] (arctic_a0002.wav by talker bdl.). The signal consists of the speech signal and the EGG signal which are simultaneously recorded.

Figure 5 shows the scatter plot of the extracted periodicity candidates just after selecting the relevant region. If it is actually relevant the user can click the “SELECT” button to proceed to the next step. The user can change the selection using the UI tools on the top left corner of the window.

Figure 6 shows the whole view of the analysis results. The top panel shows the waveform, and the middle panel shows the frequency of the candidates. The bottom panel shows the estimated SNR of the candidates. Then the user can select a region to inspect in detail. After selecting the focused region, click of the “APPLY” botton broadcasts the selected region to the other panels.

Figure 7 shows the magnified view of the selected region. The default setting uses automatic downsampling for analysis. By selecting “on” of the downsampling popup menu and clicking the “RECALCULATE” bottun starts the audio sampling rate analysis.



detailInspector

/Users/kawahara/m-file/ANNYSTRAIGHT/F0testSeries/sigslpSymposiumDec10y2018/

arctic_a0002.wav

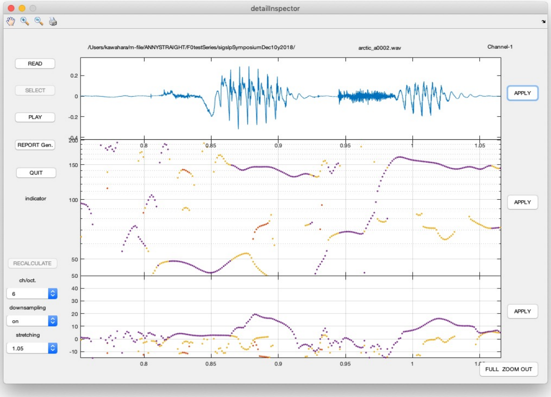
Channel-1

ch/oct. 6

downsampling on

stretching 1.05

Figure 6: *Whole view display of the analysis results. The top panel shows the waveform, and the middle panel shows the frequency of the candidates. The bottom panel shows the estimated SNR of the candidates.*



detailInspector

/Users/kawahara/m-file/ANNYSTRAIGHT/F0testSeries/sigslpSymposiumDec10y2018/

arctic_a0002.wav

Channel-1

ch/oct. 6

downsampling on

stretching 1.05

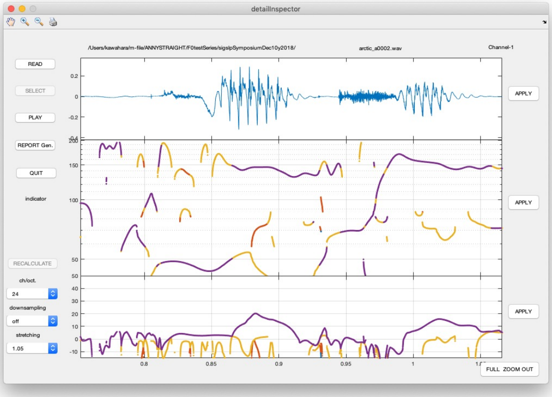
Figure 7: *Selected view display of the analysis results. The top panel shows the waveform, and the middle panel shows the frequency of the candidates. The bottom panel shows the estimated SNR of the candidates.*

Figure 8 shows the magnified view of the selected region without downsampling. In the magnified mode, by clicking the "PLAY" button, the user can playback the sound of the displayed portion.

Figure 9 shows the EGG signal analysis resluts using the similar setting. In the voiced part, both the speech and the EGG analysis results behave similarly.

7. Discussion

The core functions are general enough to be used in a wide range of applications. This implementation uses a log-linear filter setting. However, it is possible to design filter bank elements using more general frequency axes and frequency resolution [11, 12] The current release does not have event related functions, which will be available soon.



detailInspector

/Users/kawahara/m-file/ANNYSTRAIGHT/F0testSeries/sigslpSymposiumDec10y2018/

arctic_a0002.wav

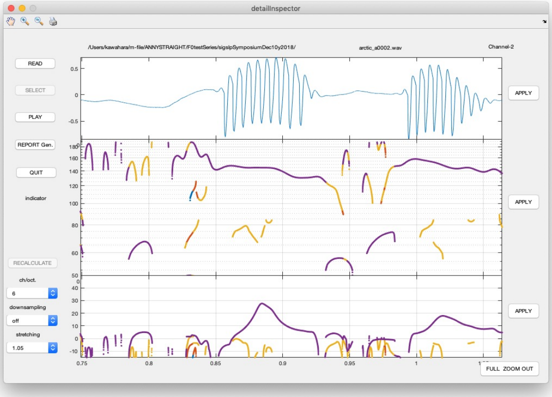
Channel-1

ch/oct. 24

downsampling off

stretching 1.05

Figure 8: *Selected view display of the analysis results without downsampling. The top panel shows the waveform, and the middle panel shows the frequency of the candidates. The bottom panel shows the estimated SNR of the candidates.*



detailInspector

/Users/kawahara/m-file/ANNYSTRAIGHT/F0testSeries/sigslpSymposiumDec10y2018/

arctic_a0002.wav

Channel-2

ch/oct. 6

downsampling off

stretching 1.05

Figure 9: *Selected view display of the analysis results without downsampling. The analyzed signal is the simultaneously recorded EGG. The top panel shows the waveform, and the middle panel shows the frequency of the candidates. The bottom panel shows the estimated SNR of the candidates.*

8. Conclusions

We introduced efficient core functions using simpler implementation of phase related attributes than Flanagan's equation. We also introduced interactive (and some are in real-time) tools by making use of this efficient implementation. The tools and constituent functions implemented using MATLAB are accessible online in GitHub. We are hoping users to acquire deeper understanding and grasp of phase related signal attributes.

9. Acknowledgements

This research was supported by KAKENHI (Grant in Aid for Scientific Research by JSPS) 16H01734 and 15H03207 and 18K00147.

10. References

[1] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, J. Kreiman, M. Kob, A. Löfqvist, S. McCoy, D. G. Miller, H. Noé, R. C. Scherer, J. R. Smith, B. H. Story, J. G. Švec, S. Ternström, and J. Wolfe, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, 2015.

[2] J. L. Flanagan and R. M. Golden, "Phase Vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, Nov 1966.

[3] Intel, "Vector Mathematics (VM): Performance and Accuracy Data," 2018, (Access date: 2018-10-12). [Online]. Available: https://software.intel.com/sites/products/documentation/

[4] H. Kawahara, H. Katayose, A. d. Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Sixth European Conference on Speech Communication and Technology*, 1999, pp. 2781–2784.

[5] H. Kawahara, K.-I. Sakakibara, M. Morise, H. Banno, and T. Toda, "A modulation property of time-frequency derivatives of filtered phase and its application to aperiodicity and fo estimation," in *Proc. Interspeech 2017*, 2017, pp. 424–428.

[6] H. Kawahara, "Pitfalls in digital signal processing," *The Journal of the Acoustical Society of Japan*, vol. 73, no. 9, pp. 592–599, 2017, [in Japanese].

[7] L. Cohen, *Time-frequency analysis: Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[8] H. Kawahara, Y. Atake, and P. Zolfaghari, "Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay." in *Icslp-2000*, Beijing, China, 2000, pp. 664—-667. [Online]. Available: http://www.isca-speech.org/archive

[9] H. Kawahara, M. M., and K. Hua, "Revisiting spectral envelope recovery from speech sounds generated by periodic excitation," in *APSIPA ASC 2018, Hawaii*. APSIPA,, 2018, pp. 1674–1683.

[10] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[11] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[12] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, Jan 2014. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17112

A. Test resuls

This appendix shows excerpts of the test results. Figure 10 shows the speed test results. The sampling frequency is 44,100 Hz, and the periodicity search range is from 55 Hz to 1000 Hz with 12 channels per octave arrangement. The temporal stretching factor is 1. The results indicate that it runs three times faster than real-time without downsampling. It also shows that it runs 12 times faster than real-time with automatic downsampling.

Figure 11 shows the standard deviation and the bias of the calculated frequency. The test signal is a periodic pulse train plus Gaussian white noise. The horizontal axis represents the local SNR and the vertical axis shows the standard deviation and the bias. The plots shows the best filter output and the surrounding filter outputs.

Figure 12 shows the periodicity measure and the input SNR. he test signal is a periodic pulse train plus Gaussian white noise. The average of measure and the SNR have close to linear relation from 5 to 45 dB SNR. The standard deviation of the measure is about 3 dB indicating that it is a reliable measure.

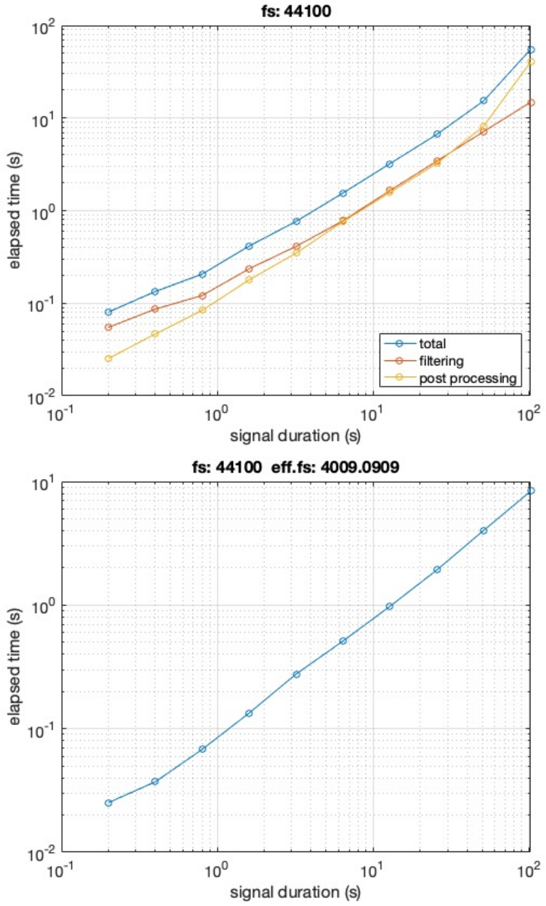


Figure 10: *Speed test results. The pper plot shows the speed without using downsampling. The lower plot shows the speed with automatic downsampling. The horizontal axis shows the duration of the test signal and the vertical axis shows the elapsed time.*

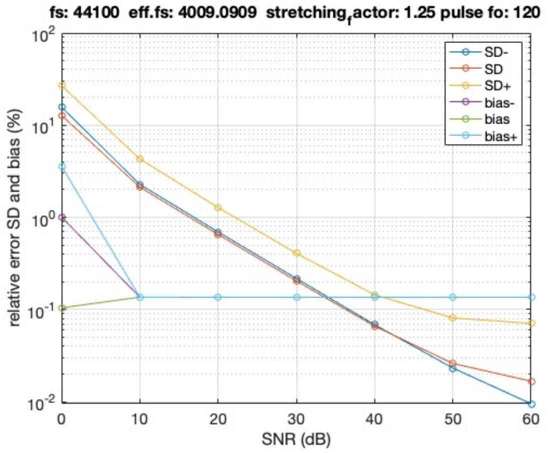


Figure 11: *Frequency estimation error and signal SNR.*

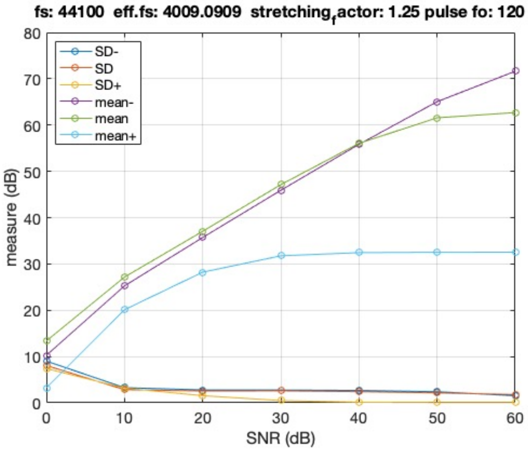


Figure 12: *Average of the periodicity measure and the standard deviation.*

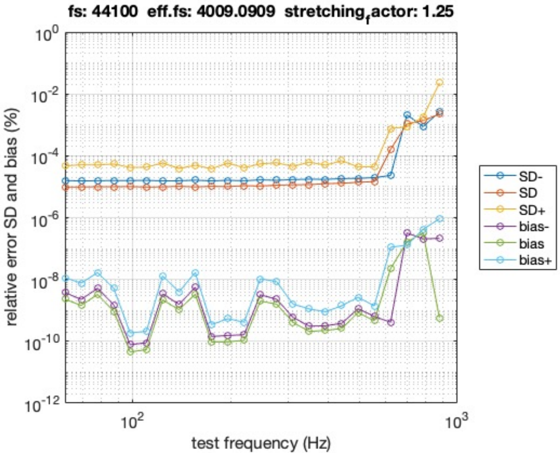


Figure 13: *The standard deviation and the bias in frequency calculation without any additive noise.*

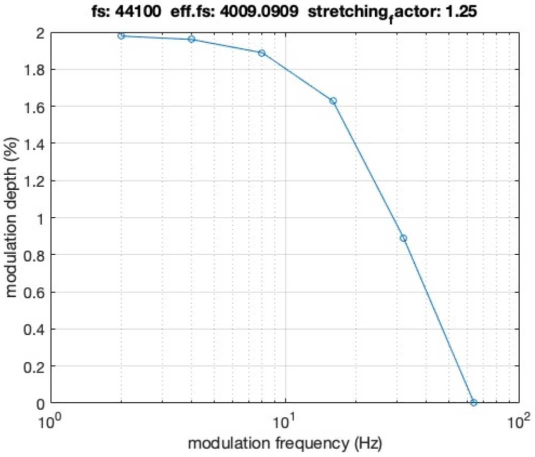


Figure 14: *Response to FM. The horizontal axis represents the modulation frequency. The modulation depth of the test signal is 2.07%.*

Figure 13 shows the standard deviation and the bias in the calculated frequency when no additive noise exists. The results indicate that they are negligible for usual situations.

Figure 14 shows the response to FM. This indicates that this procedure can track the fo trajectory of vibrato accurately.