Vol.2019-MUS-123 No.18
Vol.2019-SLP-127 No.18
2019/6/22

歌声の学習支援のための位相関連属性に基づく実時間対話的ツール

河原 英紀[1,a)]　榊原 健一[2,b)]　羽石 英里[3,c)]　萩原 かおり[3,d)]

概要：歌唱の学習では、発声に関わる多数の筋肉等を適切に制御する方法を体得しなければならず、指導者の下での訓練と長時間の練習が必要となる。ここでは、この過程を支援することを狙い、発声された歌唱音声の様々な属性を対話的にリアルタイムに可視化／可聴化して学習者にフィードバックするためのツールを紹介する。これらのツールの実装は、最近の情報デバイスの有する効率の良いマルチメディア演算とメディアの入出力機能に負っている。また、位相に関連する属性を処理に伴う副作用なく求めることができる関数と周期性による干渉を排除した信号の表現も、同様に実装の鍵となっている。本報告では、ツールの背景にある処理と実装を中心に紹介する。なお、ツールは筆頭著者の GitHub レポジトリで open source として公開している。

Kawahara Hideki[1,a)]　Sakakibara Ken-Ichi[2,b)]　Haneishi Eri[3,c)]　Hagiwara Kaori[3,d)]

1. はじめに

魅力的な歌声を得て維持するためには、適切な訓練と継続的な練習が必要になる。また、教師や言語聴覚士など日常的に声を駆使する職業では、適切な発声法の習得が発声に関する障害を避けて職業寿命を延ばすためにも重要になる。適切な発声法の習得と維持では、専門家による指導と介入が行われる。しかし、それら専門家の指示や介入を適切に理解し体得することは、初心者には容易ではない。ここでは、その過程の支援を狙い、発声に関わる音声信号の属性を対話的にリアルタイムで可視化／可聴化するツールを紹介する。ここで紹介するツールは、有用な方法を生み出すための叩き台であり、様々な場面での実際の使用経験と検証を通じて、更新して行きたいと考えている。そのため、ツールとソースコードを open source として公開している [1,2]。本報告では、ツールの技術的背景と、具体的な操作例を中心に紹介する。

[1] 和歌山大学
Wakayama University, Wakayama, Wakayama 640–8510, Japan
[2] 北海道医療大学
Health Science University of Hokkaido, Sapporo, Hokkaido 640–8510, Japan
[3] 昭和音楽大学
Showa University of Music, Kawasaki, Kanagawa 640–8510, Japan
[†1] 現在，情報処理大学
Presently with Johoshori Uniersity
[a)] kawahara@wakayama-u.ac.jp
[b)] quesokis@gmail.com
[c)] haneishi@tosei-showa-music.ac.jp
[d)] kaori@hagiwarakaori.com

2. 位相関連属性

信号の位相から求められる属性（瞬時周波数、群遅延など）を分析する際には、処理に（等価な意味で）用いられる窓関数について、振幅に関連する属性（パワーなど）を分析する際とは異なった配慮が必要になる [3]。以下では議論を簡単にするために、信号 $x(t)$ は正の周波数成分のみを含む解析信号であるとする。

2.1 瞬時周波数と群遅延

音声に含まれる成分の周波数は時間とともに変化しており、変化そのものが重要な情報を担っている。瞬時周波数 $\omega_i(t)$ は、そのように時間的に変化する信号の周波数を表現するための概念であり、位相の時間微分として定義される。位相には 2π を周期とする環状の構造がある。信号の値が複素平面上を連続的に移動する場合でも、位相には（主値を取る場合）2π 毎に不連続な飛躍が生ずる。そのため、位相を微分しようとする場合には、位相の unwrap 処理が必要とされてきた。

しかし unwrap は位相の追跡と逐次比較を含む、効率が

悪い脆弱な処理である。この問題のある unwrap を用いない次式による瞬時周波数の計算方法が 1966 年から知られている [4]。

$$\omega_i(t) = \frac{\Re[x(t)]\Im\left[\frac{dx(t)}{dt}\right] - \Re\left[\frac{dx(t)}{dt}\right]\Im[x(t)]}{|x(t)|^2}. \quad (1)$$

この式は定義から直接導かれており近似を含んでいない。瞬時周波数を求めるために用いられることのある、より簡単な Teager Kaiser Energy Operator (TKEO) は、「瞬時周波数はゆっくりと変化する」という仮定の下で成立する近似を用いて求められているため、瞬時周波数が変化する場合には、元の信号には含まれていない振動が生ずる。なお、この振動の問題と大きさの見積もりは、TKEO を提案する論文で議論されている [5]。

最近の情報デバイスではマルチメディア処理に対応するために、三角関数や複素関数などを効率よく計算する仕組みが用意されている。(例えば [6]) そのため、離散信号 $x[n]$ について、ほぼ定義通りの簡単な計算で瞬時周波数を求めることができる。

$$\omega_i[n] = \angle\left[\frac{x[n+1]}{x[n]} f_s\right], \quad (2)$$

なお、ここで f_s は、標本化周波数を表す。

群遅延 $\tau_g[k]$ は、位相の周波数微分に負号をつけたものとして定義される。離散信号／離散周波数 k での処理では、同様に次式で簡単に計算することができる。

$$\tau_g[k] = -\frac{1}{\Delta\omega}\angle\left[\frac{X[k+1]}{X[k]}\right], \quad (3)$$

2.2 解析信号とインパルス応答の包絡

インパルス応答が解析信号であるようなフィルタを通すことで、音声などの実数値の信号を解析信号にすることができる。短時間 Fourier 変換に用いられる窓関数を包絡として複素指数関数を乗ずることにより、必要なインパルス応答を用意することができる。なお、従来からよく用いられている関数 [7–9] は位相の精密な分析には不適切であり、次式による窓関数 $w_e(t; f_c, c_{\mathrm{mag}})$ を用いることとした [3, 10]。

$$w_e(t; f_c, c_{\mathrm{mag}}) = \sum_{k=0}^{K} a_k \cos\left(\frac{2\pi k f_c t}{K c_{\mathrm{mag}}}\right), \quad (4)$$

ここで c_{mag} は、フィルタの中心周波数 f_c と帯域幅との関係を調整するための係数である。次数 K を 5 とし、係数 $\{a_k\}_{k=0}^{5}$ を次の値に設定する [10]。

$$\{a_k\}_{k=0}^{5} = \{0.2624710164, 0.4265335164, 0.2250165621, 0.0726831633, 0.0125124215, 0.0007833203\} \quad (5)$$

これらを用いてインパルス応答 $w(t)$ は、次式で表される。

$$w(t) = w_e(t; f_c, c_{\mathrm{mag}}) \exp(j2\pi f_c t), \quad (6)$$

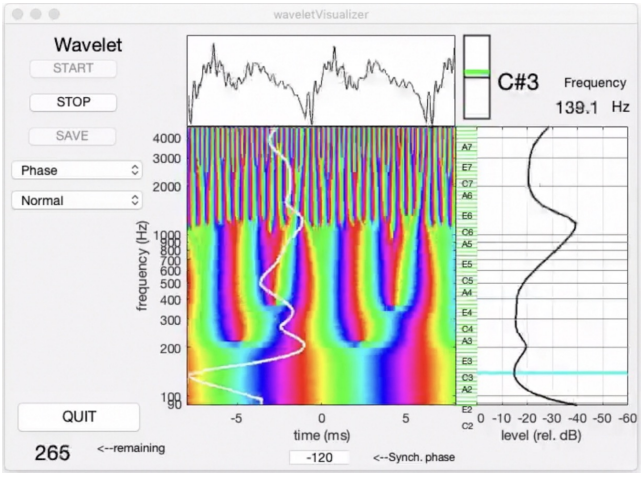

図 1 Visualization tool of a real-time wavelet analysis

2.3 基本周波数候補の選択

基本周波数 fo (fundamental frequency [11]) は、音声の重要な属性である。基本周波数の値は、話者、性別、年齢、発声内容／状況によって広い範囲で変化する。また、単一の数値として基本周波数を表すことが不適切な発声も存在する [12, 13]。ここでは、基本周波数が存在する可能性のある範囲の信号を同時に分析し、前の節で紹介した瞬時周波数と群遅延に基づいて、基本周波数の候補を選択する [14]。

中心周波数を対数周波数軸上で等間隔に配置した前述の解析信号をインパルス応答とするフィルタ群は、連続 wavelet 変換を離散周波数軸上で標本化したものに相当する。フィルタの帯域幅を、調波信号の基本波のみが単離されるように設計しておくと、フィルタ出力の瞬時周波数と群遅延は、基本波成分が卓越して含まれる場合、時間的にほとんど変化しない。この性質を利用すると、広い範囲 (SNR で 10 dB から 80dB) でフィルタ出力に含まれる正弦波成分とランダム成分の比率に比例する指標を得ることができる [14]。基本周波数という単一の数値で信号の性質を記述できない場合も、この指標を用いて複数の候補を選択することで記述できる可能性がある。

3. ツール群

こうして求められる位相関連属性をリアルタイムで対話的に表示するツール群を作成した。また、それらの属性の理解を支援するために、wavelet 変換の基礎的な出力の属性を可視化するツールと、対話的操作の過程で収集した音声資料を精密に分析する支援ツールも作成している。加えて、以前から開発している信号の可視化ツール [1] の拡張を進めている。ここでは、それらのツールを紹介する。

3.1 実時間 wavelet 分析可視化

図 1 に、wavelet 分析の可視化ツールの GUI を示す。中央のパネルの上部には、基本波の位相に同期した信号の波形が表示される。パネルの下部のイメージは、wavelet 分

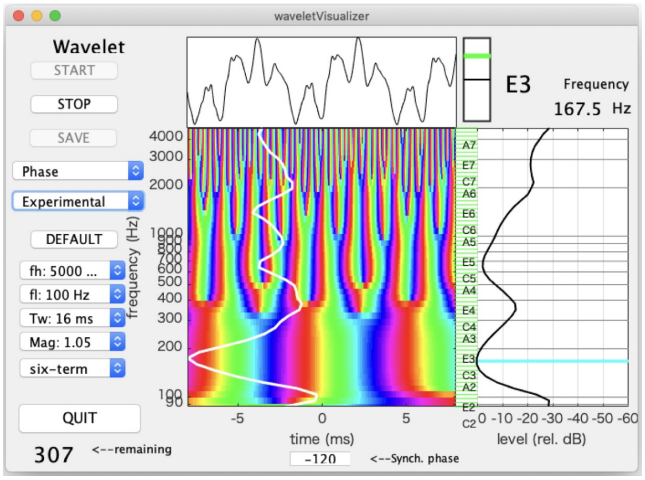

図 2　Experimental mode for detailed parameter setting.

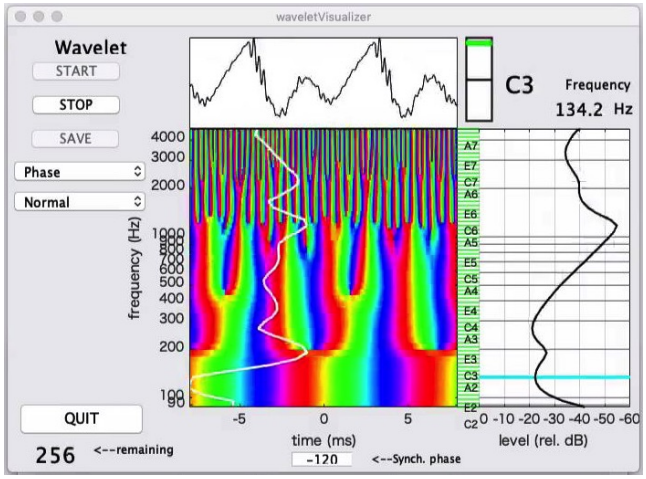

図 3　Visualization of phase of vowel /i/.

析の位相を示している。この例では、90 Hz から 4000 Hz までを、54 個の離散周波数において分析した位相を疑似カラーを用いて表示している。このパネルに重ねて表示されている白線は、各周波数における指標の値を示している。左側が高い SNR に対応している。

右側のパネルには、中央のパネルの中心位置での出力の絶対値を示す。図中の水色の水平線は、最も SNR の高い基本周波数候補の周波数を示す。中央のパネルと右側のパネルに挟まれた部分には、それぞれの基本周波数に対応する音名を記している。

右上の部分には、最も高い SNR に対応する成分の情報が表示されている。数値は、瞬時周波数、文字は最も近い音名を示す。音名の左側の矩形の中央の黒い水平線は、音名に対応する周波数を示し、矩形の上端は一半音高い周波数、矩形の下端は一半音低い周波数を示す。緑の水平線は、瞬時周波数を示している。

3.1.1　操作と操作の拡張

GUI の左側には、操作のための GUI ツールを配置した。通常は、図 1 のように必要最小限の操作を提供している。『Normal』と書かれているメニューを操作して『Experimental』を選択すると、図 2 に示すように、詳細な分析パラメタが選択できるようになる。

設定できるパラメタは、基本周波数の存在する下限と上限周波数、中央パネルが表示する時間幅、インパルス応答の包絡の伸長係数（1 以上の値で基本波成分が単離される）、包絡として用いる関数は、前述の 6 項の余弦級数、Hann 窓、Hamming 窓、Blackman 窓 [7]、Nuttall 窓 [9]、理論的に最小の時間周波数積を与える定義域が有界な関数である偏長楕円体波動関数 (PSWF, Prolate Spheroidal Wave Function。MATTLAB では DPSS という関数名)、さらに SPWF の近似である Kaiser 窓 [8] から選択できる。

なお、この拡張機能によってパラメタを設定すると、使用している計算機の処理能力が不足する場合には正常に動作しない。

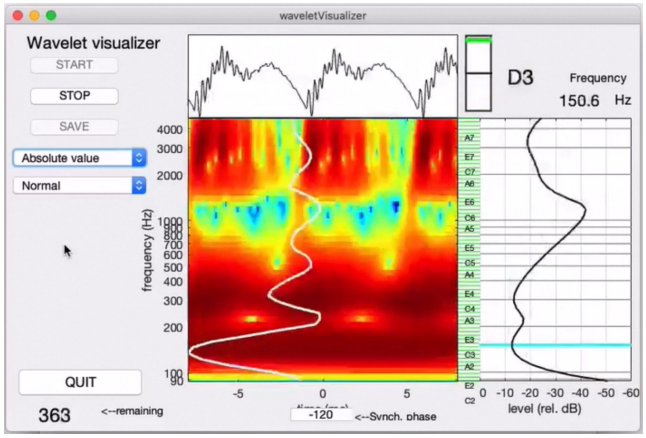

図 4　Visualization of amplitude of vowel /i/.

3.1.2　表示される属性

図 1,2 では、中央パネルには位相そのものが表示されている。このツールでは、そのほかに振幅、正規化された瞬時周波数、正規化された群遅延を選択することができる。位相情報の表示では、位相と同じ環状のトポロジーを持つ色相を位相に対応させた疑似カラー表示を用いている。その他の属性の表示では、直線状のトポロジーに適した疑似カラー表示を用いている。以下では、男性の発声した母音/i/を用いて、それぞれの表示を比較する。

3.1.2.1　位相

図 3 に男性の発声した母音/i/の位相を示す。図では-5 ms 付近と 3 ms 付近が声門閉止の時点（GCI, Glottal Closure Instant）に対応している。90 Hz から 1 kHz までの位相は、GCI に向けて集まっている。これは、GCI において、位相を揃える原因（駆動）が生じていることを示すものと解釈することができる。

3.1.2.2　振幅

図 4 に、振幅の表示例を示す。疑似カラーは、振幅の対数に対応させている。寒色が低いレベルに、暖色から茶色

Vol.2019-MUS-123 No.18
Vol.2019-SLP-127 No.18
2019/6/22

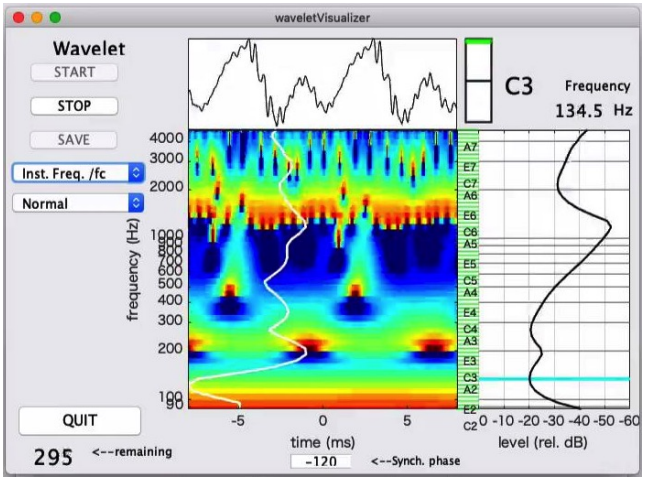

図 5　Visualization of the normalized instantaneous frequency of vowel /i/.

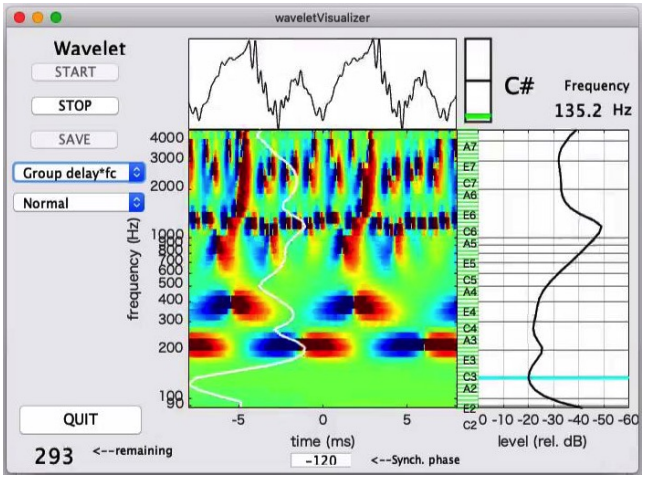

図 6　Visualization of the normalized group delay of vowel /i/.

が高いレベルに対応している。2 kHz から 4 kHz の範囲で、フォルマントに対応する声道共振が GCI（この場合は、-8 ms, 0 ms, 5 ms 付近）において周期的に駆動されている様子が明瞭に分かる。

3.1.2.3　正規化された瞬時周波数

図 5 に、それぞれのフィルタの中心周波数で正規化された出力の瞬時周波数を示す。基本周波数に相当する 134.5 Hz の付近に水平な虹色の構造が見える。これは、フィルタ出力の主要な成分が基本波である場合、出力の瞬時周波数がその成分の周波数に支配される状況を反映した結果である。同様な虹色の水平な構造は、第二フォルマントに対応する応答により、時間的変動を含んでいるが 2 kHz 付近に生じている。

3.1.2.4　正規化された群遅延

図 6 に、それぞれのフィルタの中心周波数に対応する周期を用いて正規化した群遅延を示す。ここでは、GCI に対応した時点で 600 Hz 以上で（少し湾曲しているが）縦方向の虹色の構造が見える。これは、位相の表示における位相の集束に対応する現象であり、駆動がどの時点で加えられているかを示している。

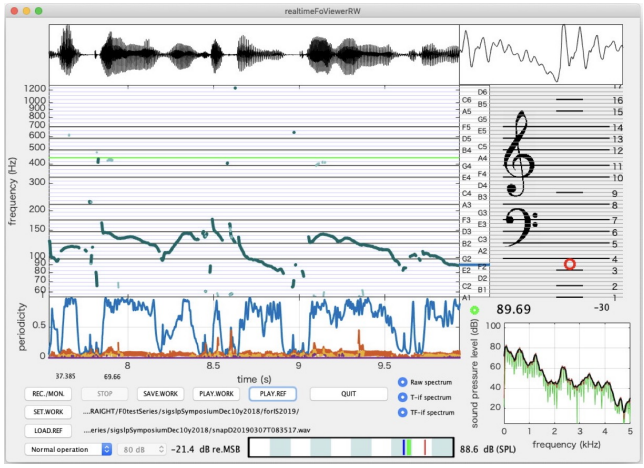

図 7　Real-time visualization of fo and spectral information with calibration and playback.

3.2　実時間 fo 可視化

前の節で紹介したツールは、技術者／開発者が位相関連属性を理解し、ツールとしてどのような属性をどのように処理するかを検討するためのものであり、学習を直接支援するためのものではない。ここで紹介する実時間の fo 可視化を中心とするツールは、学習のセッションや学習者による自習の過程を支援することを目的として開発している。併せて、（音圧が較正された状況で）音声資料を採取する機能、手本と学習者の発声を対話的に比較する機能などを用意している。

3.2.1　GUI

図 7 に、発声を表示中の GUI のスナップショットを示す。中央から左側にかけての大きな表示は、三枚のパネルから構成されており、連続的に左側に向けてスクロールし続けている。入力されあ音声を分析して得られる情報は、常にそれぞれのパネルの右端に追加され続ける。三枚のパネルは、上から順に (a) 信号の波形、(b)fo の候補、(c) 周期性の程度、を表している。fo の候補を表示するパネルの縦軸は、周波数を周波数の対数に比例する位置に配置している。このパネルには、多数の淡い色の横線が表示されている。これらは、西洋音楽で用いられる半音階に対応している。また、黒い水平線は、低音部譜表と高音部譜表の五線譜に対応している。

この中央の段の右側には、音名を表示した縦長の部分と、さらに右側には五線譜が表示されている。入力信号のピッチが明瞭な場合には、基本周波数に対応する位置に赤色の全音符が表示される。また、ピッチが明瞭な場合には、これらの下に、緑色の円の表示と、基本周波数の値（単位：Hz）が表示される。

右上のパネルは、基本波の位相に同期して、入力信号の波形を拡大表示している、右下のパネルは、求められた基

本周波数の情報を利用して、(a) パワースペクトル、(b) 駆動の周期性に基いて生ずる時間方向の変動を抑圧したパワースペクトル、(c)(b) から更に周波数方向の変動を抑圧したスペクトル包絡 [15] を表示している。

三枚のパネルの最下段は、フィルタ出力の周期性の程度を 0 から 1 までの数値で表している。1 が完全に周期的な信号の場合に相当する。なお、この周期性を表す指標は、フィルタ出力から求めらる SNR に基づいて、直感的に把握しやすい範囲の数値に変換されたものである。

下側の操作パネルには、分析の停止／再開や、停止の直前の記録内容の再生、手本となるファイルの読み込み、手本の再生、音圧レベルの較正のためのボタンが用意されている。次の節では、音圧レベルの較正について説明する。

3.2.2 音圧レベルの較正

GUI の最下段には、横長のレベルメータが表示されている。レベルメータの右端は、D/A 変換で得られる最大の瞬時値に対応している。メータの中の赤線は、瞬時値、黒線はパワー（自乗平均値）、緑線は、時間的に平滑化されたパワーを表す、このメーターを用いて、マイクの位置での音圧が目標とする値（この例では 80 dB）となった時に、較正情報を取得するためのボタンをクリックする。すると、以降の表示と録音では、較正された音圧レベルがスペクトルの表示で用いられ、録音音声をファイルに記録する際に、較正に必要な情報が、ファイルのヘッダに記録される。

4. 既存の実時間対話的ツール SparkNG との連携と拡張

これらのツールのために用意された分析用の関数を利用することにより、音声からの声道形状の実時間表示、発声シミュレータのパラメタを実際の音声から求める機能などを追加することで拡張する予定である。

5. おわりに

ここでは、位相に関連する属性（瞬時周波数および群遅延）が高速かつ高精度に計算できる状況を利用して開発した、対話的にリアルタイムで情報を可視化／可聴化するツールを紹介した。対話的にリアルタイムで豊かなフィードバックをこれらのツールにより提供することで、適切な発声法の習得や維持の過程を支援したいと考えている。現在、これらは MATLAB により開発されソースコードなどが GitHub を利用して open source として公開している。併せて,MATLAB を必要としないコンパイル済みのアプリケーションも用意している。ぜひ、様々な状況で試用し、経験をフィードバックして頂きたい。なお、今後は、より広く応用できるように開発環境および対象とする情報デバイスについても検討を進める予定である。

謝辞

本研究は科研費基盤 (A)15H03207、基盤 (B)16H01734、基盤 (C)18K00147 の支援を受けている。

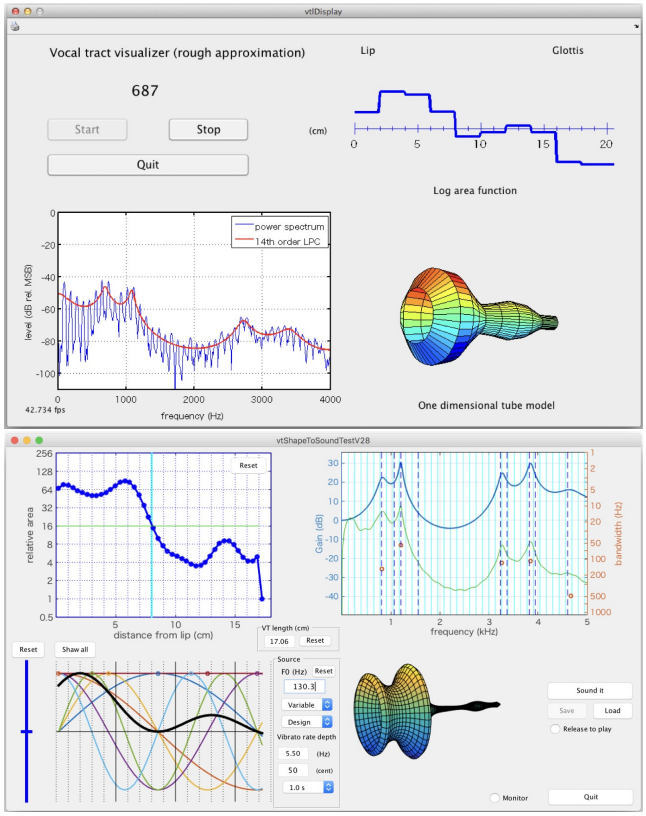

図 8 Vocal tract visualizer and voice production simulator.

参考文献

[1] Kawahara, H.: SparkNG MATLAB real-time/interactive tools for speech science research and education, GitHub (online), available from ⟨https://github.com/HidekiKawahara/SparkNG⟩ (accessed Last access: 2019-05-29).

[2] Kawahara, H.: Analytic signal-based source information analysis for YANGstraight and real-time interactive tools, GitHub (online), available from ⟨https://github.com/HidekiKawahara/YANGstraight_source⟩ (accessed 2019-05-29).

[3] 河原英紀：ディジタル信号処理の落とし穴，日本音響学会誌， Vol. 72, No. 9, pp. 592–599 (2017).

[4] Flanagan, J. L. and Golden, R. M.: Phase Vocoder, *Bell System Technical Journal*, Vol. 45, No. 9, pp. 1493–1509 (online), DOI: 10.1002/j.1538-7305.1966.tb01706.x (1966).

[5] Maragos, P., Kaiser, J. F. and Quatieri, T. F.: Energy separation in signal modulations with application to speech analysis, *IEEE transactions on signal processing*, Vol. 41, No. 10, pp. 3024–3051 (1993).

[6] Intel: Vector Mathematics (VM): Performance and Accuracy Data, Intel® Math Kernel Library 2018 (online), available from ⟨https://software.intel.com/sites/products/documentation/⟩ (accessed 2018-10-12).

[7] Harris, F. J.: On the use of windows for harmonic analysis with the discrete Fourier transform, *Proceedings of the IEEE*, Vol. 66, No. 1, pp. 51–83 (1978).

[8] Kaiser, J. and Schafer, R. W.: On the use of the I_0-sinh

window for spectrum analysis, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 28, No. 1, pp. 105–107 (1980).

[9] Nuttall, A. H.: Some windows with very good sidelobe behavior, *IEEE Trans. Audio Speech and Signal Processing*, Vol. 29, No. 1, pp. 84–91 (1981).

[10] Kawahara, H., Sakakibara, K.-I., Morise, M., Banno, H., Toda, T. and Irino, T.: A new cosine series antialiasing function and its application to aliasing-free glottal source models for speech and singing synthesis, *Proc. Interspeech 2017*, Stocholm, pp. 1358–1362 (2017).

[11] Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., Švec, J. G., Ternström, S. and Wolfe, J.: 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 (online), DOI: 10.1121/1.4919349 (2015).

[12] 榊原健一: 発声と声帯振動の基礎 (やさしい解説), 日本音響学会誌, Vol. 71, No. 2, pp. 73–79 (2015).

[13] 榊原健一: 世界の歌唱法 : 様々な歌唱様式における supranormal な声 (< 小特集 > 歌声の科学), 日本音響学会誌, Vol. 70, No. 9, pp. 499–505 (2014).

[14] 河原英紀, 榊原健一, 森勢将雅, 石本祐一: 偏長楕円体波動関数を包絡とする解析信号によるオーディオ標本化周波数での実時間 fo 候補抽出について, 信学会音声研究会, Vol. 118, No. 497, SP2018-113, pp. 305–310 (2019).

[15] Kawahara, H., Morise, M. and Hua, K.: Revisiting spectral envelope recovery from speech sounds generated by periodic excitation, *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, pp. 1674–1683 (2018).