一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2022-2(2022-06)

音声分析合成基盤 WORLD の GUI 実装と見えてきた課題

河原　英紀†　　森勢　将雅††

† 和歌山大学　〒640-8510 和歌山市栄谷 930
†† 明治大学　〒164-8525 東京都中野区中野 4-21-1
E-mail: †kawahara@wakayama-u.ac.jp, ††mmorise@meiji.ac.jp

あらまし　音声分析合成基盤である WORLD vocoder に、基本的な分析合成の支援、対話的なパラメタ操作による知覚的影響の確認などを容易にする GUI を MATLAB を用いて開発している。この過程で得られた知見と、将来の課題について議論したい。

キーワード　ピッチ抽出器、非周期性指標、スペクトル包絡、パラメタ操作、モーフィング、グラフィカルユーザインタフェース

Issues emerged from implementation of GUI tools for WORLD VOCODER

Hideki KAWAHARA† and Masanori MORISE††

† Wakayama University,　930 Sakaedani, Wakayama, 640-8510, Japan
†† Meiji University,　4-21-1 Nakano, Nakano-ku, Tokyo, 164-8525 Japan
E-mail: †kawahara@wakayama-u.ac.jp, ††mmorise@meiji.ac.jp

Abstract　We discuss issues emerged from development of graphical user interfaces (GUIs) for the WORLD VOCODER. The GUIs consist of the fundamental analysis and synthesis handler, the interactive parameter manipulation and resynthesis tool, and the interactive assisting tool for assigning landmarks for voice morphing. We also describe helpful features provided in software development environment prepared for MATLAB.

Key words　pitch extractor, aperiodicity index, spectral envelope, parameter manipulation, morphing, graphical user interface

1. はじめに

深層学習に基づく音声応用技術の発展はめざましく [1]、状況によっては人間を凌ぐ能力も実現されるようになっている。しかし、人間がどのように音声を生成し（言語情報に限らずパラ言語・非言語情報も含めて）認識しているかについての理解には結びついていないように思える。ここでは、従来の信号処理技術にもとづく表現ではあるが、対話的に分析・操作・再合成できる環境を用意することで、人間による音声生成・知覚・認識の理解を支援する可能性を議論したい。

2. 背　　景

筆頭著者は、1986 年に開設された国際電気通信研究所（ATR）において、Symbolics の Lisp machine の上に実装された対話的音声研究環境 SPIRE [2] に出会って強い印象を受けた。自分のために同様な対話的音声研究環境が欲しくなり、Lisp machine と比較すると圧倒的に非力であるが、当時普及していた PC9801 という Z80 ベースのマシンの上に Turbo Pascal を用いて音声知覚の研究のための環境を実装した [3][1]。

10 年ほどの間、この対話的環境への関心がその先に進むことはなかった。しかし、1997 年の legacy-STRAIGHT の発明 [4] と効率の良い開発環境である MATLAB により、より自由度の高い操作が可能となり、音声の対話的研究環境への関心が再起動した。STRAIGHT の音声パラメタの操作用の GUI の提供から始まり、二つの音声間のモーフィングが可能となった [5]。その音声モーフィングは数段階を経て、時変多属性任意事例数モーフィングへと一般化された [6]〜[8]。

MATLAB による対話的実時間ユーザインタフェースの開発は、2016 年に新しい開発環境である Appdesigner が導入され、実時間処理も Audiotoolbox が導入されたことにより、飛躍的に容易になった。概算であるが、人間が記述すべきコードの量は 50 分の 1 程度になっている [9]。マシンの計算能力も、筆頭著

(注1)：FFT をインラインアセンブラで実装したり、当時の 16 色のみのカラーディスプレイでディザパターンを設計することで、256 色の擬似カラーでスペクトログラムを表示させていた。最終的に 25000 行のコードになった。このソフトウェアは、1989 年に NTT アドバンステクノロジ（株）から『音声工房』として発売された

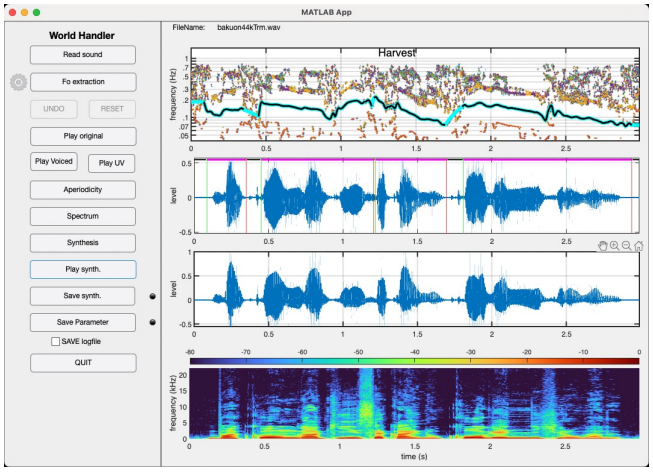


図 1　GUI of the basic analysis and synthesis.

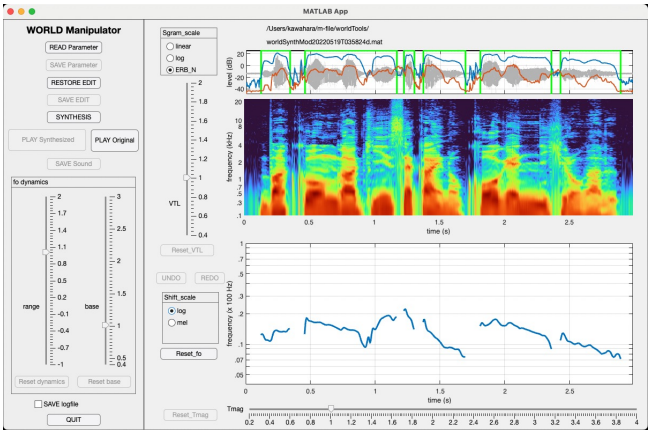


図 2　GUI of interactive parameter manipulation and synthesis.

者が本格的に計算機を使い始めた 1974 年頃と比較すると、2018 年頃には 1 億倍になっているとの試算もある。さらに、アプリケーションの開発で、アプリケーションを実行中にコードを書き換えて支障なく実行に反映されるなど、開発環境の進歩も著しい [10]。

2013 年に帰国前日にあわてて実装したモーフィング用の資料を準備するツールは、Appdesigner 以前の開発環境である GUIDE を用いており、MATLAB により用意されている対話用の基盤と整合しない部分が多く、非常に使いにくいものになってしまっていた。legacy-STRAIGHT、TANDEM-STRAIGHT と続けて開発してきた VOCODER も、間に合わせの実装が混在するようになり保守が困難になりつつある。幸い、これらの STRAIGHT と WORLD は概念が類似しており、WORLD にはオープンソースで複数の言語の実装が作られるなどの将来につながる利点・可能性がある [11], [12]。モーフィングや物理パラメタ操作と再合成用に STRAIGHT の上で開発してきたツールも、概念レベルでは WORLD と共通している。そこで、この機会に、WORLD を基盤として最近の開発環境の上に一連のツール群を構築することとした。以下、開発を開始したツールの現状を紹介し、見えてきた課題などについて議論していきたい。

3.　実装したツール

用意したツールは、3 種類である。worldHandler は、基本的な音声の分析と合成を行う。worldManipulator は、分析された音声のパラメタを変形して再合成する。morphingAligner は、拡張されたモーフィングに必要となる付加情報の作成を支援する。

3.1　worldHandler

図 1 に worldHandler の GUI の例を示す。左側のパネルには、操作用のボタンなどが配置されている。右側のパネルは、情報の可視化用である。

左側パネルの最上段の Read Sound ボタンのクリックにより、音声ファイルを選択して読み込む。MATLAB の音声ファイル読み込み用の関数がサポートしている全ての種類のファイルを読み込むようにしている。

次の段の Fo extraction ボタンのクリックにより、基本周波数が求められる。WORLD のピッチ抽出器である Harvest は、音声分析合成の際の品質劣化を避けることを狙い、有声音と判断する割合が高く設定されている。目的によっては、この偏りは不都合であるため、この GUI には、有声／無声境界を対話的に修正する機能を加えてある。また、求められた基本周波数の軌跡を手作業によって修正することも可能にしている。このボタンの直下の UNDO と REDO は、間違いの取り消しとやり直しのために用意した。その下の Play original と、さらにその下の Play Voiced と PlayUV は、有声／無声境界を試聴で確認するために用意した。

その下の Aperiodicity ボタンをクリックすると、設定された基本周波数と有声／無声境界に基づいて、非周期性指標が計算される。同時に、その下の Spectrum ボタンが使用可能になる。Spectrum ボタンのクリックによりスペクトル包絡が求められ、その下の Synthesis ボタンが使用可能になる。

Synthesis ボタンのクリックにより合成音声が作成され、以下のボタンで再生し、分析結果をファイルとして格納することができるようになる。

MATLAB のグラフには、拡大や移動、読み出しの機能が既定値として用意されている。この GUI でもそれらの機能を利用できる。また、MATLAB のグラフ間の軸の連動機能を設定してあり、基本周波数、波形、スペクトル包絡のいずれのグラフで時間軸に対して行った操作も他のグラフを連動して変化させる。なお、原音声も合成音声も、グラフが拡大表示されている場合には、グラフとして見えている部分に対応する音だけが再生される。

3.2　worldManipulator

図 2 に worldManipulator の GUI を示す。操作項目が多いため、操作用のツールが右側の可視化部分にも置かれている。

左の最上段の READ Parameter ボタンにより、前の節で紹介した worldHandler により分析され格納されたパラメタを読み込む。読み込まれたパラメタは、右側のパネルのグラフに可視化される。最上段は、波形とパワーの系列を示している。この表示とその下のスペクトログラムの表示、さらに最下部の基本周波数の表示は、MATLAB の機能により時間軸が連動しており、拡大・移動などの操作は、全ての表示に即座に反映される。

スペクトログラムと基本周波数の周波数軸は、工学的な対数軸に加え、内耳の基底膜上の周波数配置と対応する ERB_N rete 軸 [13], [14]、音の高さの知覚に対応する Mel 軸が用意されている。

基本周波数の操作には、左のパネルにある二つのスライダーを使う方法と、右の基本周波数の可視化上に直接手書きする手段を用意した。スライダーでは、可視化された表示を上下に平行移動する右側のスライダーと、可視化された表現上の変動を定数（負の数値も含む）倍する手段を用意した。基本周波数の背景にある生理学的な構造が示唆する対数軸上での平行移動は、知覚的には一定の印象を与えないことが示唆されている。それらが、実際にはどのような印象となるかを、この GUI のスライダーの操作で、対話的に確認することができる。なお、手書きによる操作には、取り消しと再試行のための UNDO、REDO を用意した。

可視化パネルの上部にあるスライダーは、声道長（VTL: Vocal Tract Length）の比例的伸縮を行う。スライダー操作終了と同時に声道長変化によるスペクトログラムの変化が可視化される。

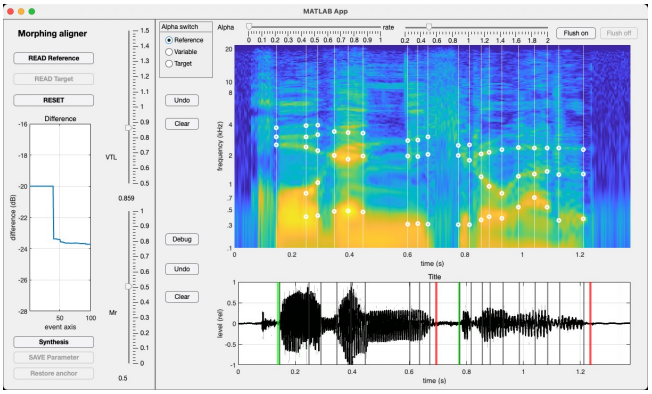
可視化パネルの最下段にあるスライダーは、時間軸の比例的伸縮を行う。スライダーの操作は、即座に可視化に反映される。この場合、可視化されたイメージは固定して表示され、軸のメモリが比例的に変化する。

対話的に試行錯誤を繰り返している状態で、途中の操作状況を保存したい場合には SAVE EDIT ボタンを用い、最終的な結果と操作されて変化したパラメタを保存したい場合には SAVE Parameter を用いる。

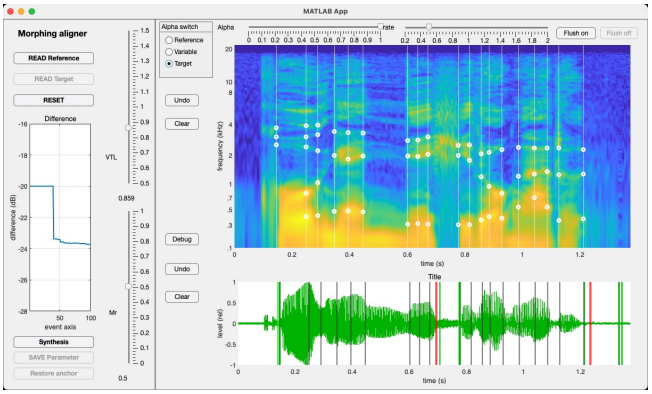
3.3 morphingAligner

図 3 に、二つの事例間の対応の調整を支援するツールである morphingAligner の GUI を示す。これらのスナップショットは、対応が調整された後の状態を示している。(a) が、固定して表示されるスペクトログラムと、その上に配置された目印を示す。この例では男性による音声を固定されるものとして用いた。ここでは Reference と名づけている。(b) は、時間軸と時間周波数上の点を操作して変形されたスペクトログラムを示している。この例では女性による音声を変形されるものとして用いた。ここでは、Target と名づけている。(c) は、スペクトログラムの透明度を調整して、同じ透明度で重ねたものである。目印として設置した白線と白円の位置が同一であることが分かる。

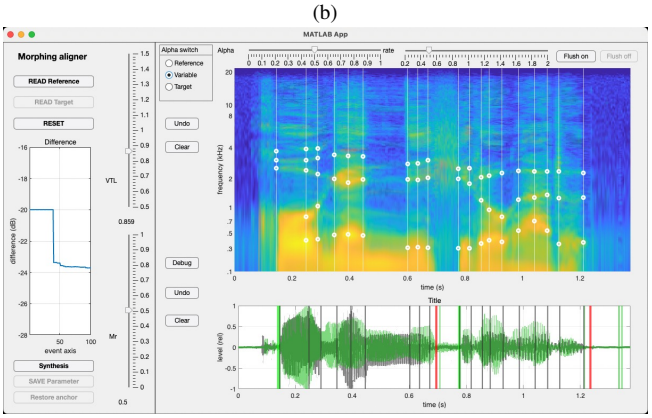
実際の操作は、Reference と Target のパラメタを読み込んで、同じ透明度で重ねた状態で、時間方向の目印を逐次的に増やしながら位置を調整し、左側のグラフで示されている二つのスペクトログラムの距離が減少するようにすることを繰り返すことから始める。次に、位置の調整が終わったところで、両方の音声の背景にある声道長の違いを左上のスライダーによって調整する。こうして、時間軸と声道長比の調整が終わったところで、周波数方向の細かな違いを、目印を設定し位置を調整することで、さらに二つのスペクトログラムの違いが減少するように調整する。この調整結果をファイルに記録し、読み込んで再開するためのボタンを用意している。Synthesis ボタンと、Mr と書かれたスライダーにより、整列された二つの事例のパラメタを、



(a)



(b)



(c)

図 3　GUI of interactive tool for assigning corresponding points. (a) Reference voice spoken by a male, (b) target voice spoken by a female, and (c) mixed visualization.

スライダーに示された比率で混合して、モーフィング音声を合成する。

二事例間のモーフィングの場合、割合を 0.5 とした場合が、二つの事例のパラメタの整列の悪さの影響が最も大きく現れる。そのため、適切な整列の支援を目的とするこのツールでは、全てのパラメタを同一の混合比で設定している。

拡張されたモーフィングには、このような制約はない。それぞれの音声パラメタの時刻毎にそれぞれ異なった混合割合を設定することが可能なように構成されている。また、そのための合成関数も用意している。問題は、設定の自由度が大きいため、どのように操作することができるようにするか、適切な概念を設定しツールとして実装することが難しいことにある。

4. 課題など

MATLAB のグレフィックスが提供する透明度の操作を利用することで、非常に困難であった手作業による音声事例間の対応づけが、直感的に実行しやすい操作となった。この GUI で用いた方法を応用すれば、worldManipulator での操作に、時間反転を含む非線形なものを追加して自由度を増すことができる。

本資料では、従来型の VOCODER のパラメタを対話的に操作するツールを紹介した。このような既存の知識に（ある意味では）縛られたパラメタ／情報表現の上での操作にとどまらず、深層学習を通じて明らかになる潜在変数に基づく操作ツールへと拡張することが、音声の生成と知覚の深い理解につながる可能性があるようにも思う。

5. まとめ

音声分析合成基盤である WORLD vocoder に、基本的な分析合成の支援、対話的なパラメタ操作による知覚的影響の確認などを容易にする GUI を MATLAB を用いて開発している。この過程で得られた知見と、将来の課題について議論したい。ここで紹介したツールのソースと、ツールの操作法を紹介するムービーなどの資料は筆頭著者の GitHub リポジトリにリンクしておく。

謝　辞

本研究は科研費 18K00147, 18K10708, 19K21618, 21H04900 の支援を受けた。

文　献

[1] 全　炳河，“深層学習によるテキスト音声合成の飛躍的発展，” 電子情報通信学会誌，vol.105，no.5，pp.413–417，2022.

[2] V. Zue, D. Cyphers, R. Kassel, D. Kaufman, H. Leung, M. Randolph, S. Seneff, J. Unverferth, and T. Wilson, “The development of the mit lisp-machine based speech research workstation,” ICASSP ’86. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.11, pp.329–332, 1986.

[3] 河原英紀，“音声知覚過程研究支援環境のユーザインタフェース，” 1987．聴覚研究会資料, H-87-32,.

[4] H. Kawahara, I. Masuda-Katsuse, and A. deCheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,” Speech Communication, vol.27, no.3-4, pp.187–207, 1999.

[5] H. Kawahara and H. Matsui, “Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation,” 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)., vol.1, pp.I–I, 2003.

[6] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, “Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown,” 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.3905–3908, 2009.

[7] H. Kawahara, M. Morise, H. Banno, and V.G. Skuk, “Temporally variable multi-aspect n-way morphing based on interference-free speech representations,” 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp.1–10, 2013.

[8] H. Kawahara and V. Skuk, “Voice morphing,” Oxford Handbook of Voice Perception, eds. by S. Frühholz and P. Belin, pp.683–706, Oxford University Press, Oxford UK, 2018.

[9] 河原英紀, 小谷野進司, 亀川　徹, 丸井淳史, “音響アプリケーションのユーザインタフェース開発事例,” 日本音響学会誌，vol.77，no.4，pp.239–247，2021.

[10] 加藤　淳，“インタフェース・デザインの勘所,” 日本音響学会誌，vol.77，no.4，pp.231–238，2021.

[11] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Information and Systems, vol.99, no.7, pp.1877–1884, 2016.

[12] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” Proc. Interspeech, pp.2321–2325, 2017.

[13] B.C.J. Moore and B.R. Glasberg, “Suggested formulae for calculating auditory‐filter bandwidths and excitation patterns,” The Journal of the Acoustical Society of America，vol.74，no.3，pp.750–753，1983. https://doi.org/10.1121/1.389861

[14] D.D. Greenwood, “A cochlear frequency-position function for several species—29 years later,” The Journal of the Acoustical Society of America，vol.87，no.6，pp.2592–2605，1990.