Vol.2019-MUS-123 No.55
Vol.2019-SLP-127 No.55
2019/6/23

リアルタイム音声分析合成エフェクター CHERRY PIEの実装

寺島 涼[1,a)]

概要：VOCODER 方式の音声分析合成システムは，ピッチ操作やスペクトル変形に対する柔軟性が高く，音声研究では広く用いられている．しかし，予め高度な解析処理を必要とするため，分析から合成までの全てを低遅延でリアルタイムに処理することが難しく，実際の楽曲制作やライブなどで実用的に広く利用されるまでには至っていない．制作現場での実用性を高めるためには，品質や分析精度の追及よりも，比較的高い品質で，かつ，頑健性が高く，容易に扱えるという観点が特に重要になると考えられる．筆者らは，VOCODER 方式の音声分析合成特有の高い柔軟性を保持しつつ，実用的な品質と頑健性を有し，分析から合成までをリアルタイム，かつ，低遅延で処理するというコンセプトを基に，音声分析合成エフェクター CHERRY PIE を開発した．本稿では，リアルタイム音声分析合成を実現するためのアルゴリズムの骨格，及び，考え方について述べる．

Implementation of Real-Time Speech Analysis Synthesis Effector CHERRY PIE

Ryo Terashima[1,a)]

1. はじめに

VOCODER 方式の音声分析合成システムは，ピッチ操作やスペクトル変形に対する柔軟性が高く，聴覚実験や音声合成システムのコア部分として，音声研究では広く用いられている [1], [2], [3], [4], [5], [6], [7]．しかし，それらの多くは，予め高度な解析を行う必要があるため，フルバンドの音声において，分析から合成までの全てを低遅延でリアルタイムに処理することが難しい．リアルタイム動作を目的とした VOCODER 方式の音声分析合成システム [8] の取り組みも既に存在するが，実際の楽曲制作やライブなどで実用的に広く利用されるまでには至っていない．制作現場での実用性を高めるためには，品質や分析精度の追及よりも，比較的高い品質で，かつ，頑健性が高く，容易に扱えるという観点が特に重要になると考えられる．

そこで筆者らは，VOCODER 方式の音声分析合成特有の高い柔軟性を保持しつつ，実用的な品質と頑健性を有し，分析から合成までをリアルタイム，かつ，低遅延で処理するというコンセプトを基に，音声分析合成エフェクター CHERRY PIE を開発した [9], [10]．CHERRY PIE は，VST (Virtual Studio Technology) 及び，AU (Audio Units) のオーディオプラグイン形式で実装されており，DAW (Digital Audio Workstation) 等の楽曲制作ソフトウェアから，簡単に利用することが可能となっている．CHERRY PIE には，スラングで「簡単にできること」という意味があり，高度な音声分析合成技術を簡単に扱えるようにしたいという思想から，本システムの名前に採用した．

本稿では，リアルタイム音声分析合成を実現するためのアルゴリズムの骨格，及び，考え方について述べる．更に，CHERRY PIE 上に搭載されている各エフェクトについても簡単に紹介する．

[1] クリプトン・フューチャー・メディア株式会社
CRYPTON FUTURE MEDIA, INC
[a)] terashima@crypton.co.jp

Vol.2019-MUS-123 No.55
Vol.2019-SLP-127 No.55
2019/6/23

2. 音声分析合成アルゴリズム

2.1 処理概要

音声分析合成の処理概要を**図 1** に示す．入力音声に対して，基本周波数（F0）推定を行い，更に，F0 エフェクト処理が施された合成 F0 を作成する．次に，分析 F0 と合成 F0 に基づき，分析ピッチマークと合成ピッチマークを配置する．分析ピッチマークごとにスペクトル包絡・非周期性指標を推定し，それぞれにエフェクト処理を行う．続いて，スペクトル包絡・非周期性指標に基づき，周期成分・非周期成分を算出し，それぞれにエフェクト処理を行う．周期成分・非周期成分より，最小位相応答で得られた単位波形を合成ピッチマーク上に重畳加算することで合成音声を出力する．各エフェクト処理が特に行われない場合は，加工のされていない分析合成音声が出力される．

分析，及び，合成処理では，STFT (Short-Time Fourier Transform) に基づく処理が行われる．その際の窓幅を W, FFT (Fast Fourier Transform) 点数を N，フレームシフト幅を S で表す．なお，フレームシフト幅は F0 推定でのみ用いられる．FFT 点数は，各処理に応じて変更することで効率化を図ることも可能であるが，簡単のため，本稿では同じ値を用いる．

実用上は，サンプリング周波数が 44,100 Hz の場合において，下記の (1), (2), (3) いずれかの設定を用いることが多い．

（1）W=1024，N=1024，S=256
（2）W=1536，N=2048，S=256
（3）W=2048，N=2048，S=256

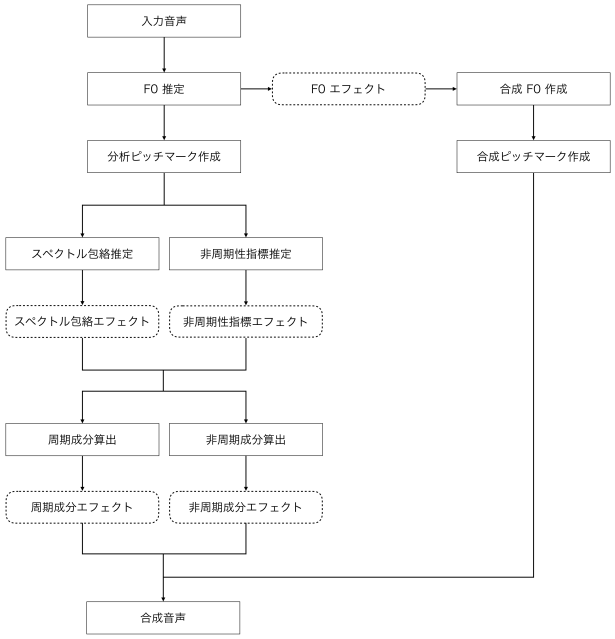

図 1 音声分析合成の処理概要

それぞれの設定におけるアルゴリズム遅延はおよそ, 23msec, 35msec, 46msec となる．(1) は，アルゴリズム遅延を極めて低くすることが可能だが，低い F0 に対する分析精度が不十分になる．(2), (3) と窓幅 W を広げるに従って，アルゴリズム遅延が増大するが，低い F0 に対する分析精度は向上する．

CHERRY PIE では設定画面から，これらの数値を変更することができる．入力音声の特性や使用用途に応じて，柔軟に対応することが可能である．なお，公開されている CHERRY PIE のデモムービー [10] では，(1) の設定を利用している．また，若干の音質への影響を伴うが，アルゴリズム遅延をより低減する方式に向けた検討を進めており，今後更に取り組む予定である．

2.2 分析アルゴリズム

2.2.1 振幅スペクトルに対する包絡成分・微細構造の分離

本手法では，STFT で得られた振幅スペクトルより，ラグ窓 [11] を用いて包絡成分と微細構造に分離する．この分離手法は，後節で述べる F0 推定，及び，非周期性指標推定で用いられる．ラグ窓は式 (1) で表される．L は窓長であり，短いほど平滑化された包絡成分が得られる．

$$\omega_\tau = \frac{(L!)^2}{(L+\tau)!(L-\tau)!} \tag{1}$$

本手法では，下記の (a), (b), (c) のいずれかを算出し，ラグ窓をかけた上で再度フーリエ変換を行い包絡成分を求める．

（a）振幅スペクトルの逆フーリエ変換
（b）パワースペクトルの逆フーリエ変換
（c）対数パワースペクトルの逆フーリエ変換

その際，(b) または (c) を用いて包絡成分を得た場合は，元の振幅スペクトルと同次元の線形振幅に戻す．(a), (b), (c) の選択によって，得られる包絡成分の傾向が異なり，本手法で利用する際は処理に応じて選択を変えている．対数振幅上で見ると，(c) では，振幅スペクトルの微細構造の山谷の中心を辿る包絡成分が得られ，(b)，(a) の順で，より山の頂点に近い包絡成分が得られる．

続いて，得られた包絡成分で元の振幅スペクトルを除することにより微細構造を算出する．すなわち，(a), (b), (c) いずれの場合も，分離された包絡成分と微細構造の積が元の振幅スペクトルと一致するように算出される．

振幅スペクトルに対する包絡成分・微細構造の分離の例を**図 2** に示す．この例では，(a) を用いている．

2.2.2 F0 推定と分析ピッチマーク作成

F0 は，フレームシフト幅 S ごとに推定される．まず，窓幅 W，FFT 点数 N の STFT に基づき振幅スペクトルを算出し，2.2.1 の (a) を用いて包絡成分と微細構造に分離する．更に，2.2.1 の (c) を用いて，微細構造を再度分離することで，対数振幅上において，山谷の中心が概ね 0 dB 付

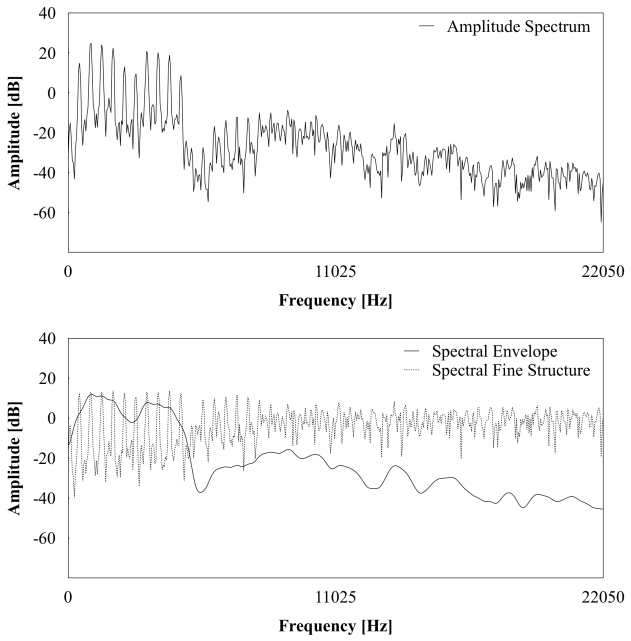

図 2 振幅スペクトル（上），分離された包絡成分・微細構造（下）

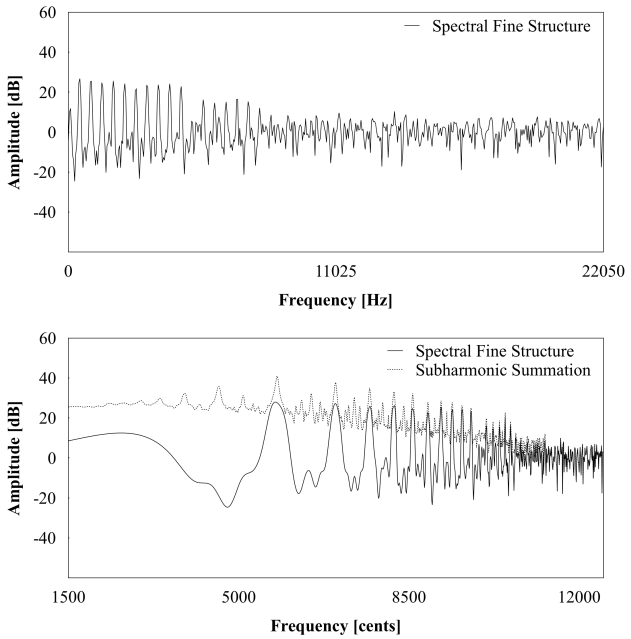

図 3 微細構造（上），スプライン補間された微細構造と SHS （下）

近を辿るように補正された微細構造を獲得する（**図 3** 上）．

この補正された微細構造を対数振幅上でスプライン補間し，対数周波数軸上で一定間隔の微細構造を得て，SHS (Subharmonic Summation) を算出する（図 3 下）．そして，SHS のピークを推定 F0 候補とする．

スプライン補間と SHS に関するアーキテクチャは，Ikemiya ら [12] の手法を参考にした．筆者らの手法では，フォルマントの影響の少ない微細構造に対し SHS を適用することで，F0 推定の精度と頑健性の向上を図っている．筆者らは，この F0 推定手法を，SFS-SHS (Spectral Fine Structure Subharmonic Summation) と呼んでいる．

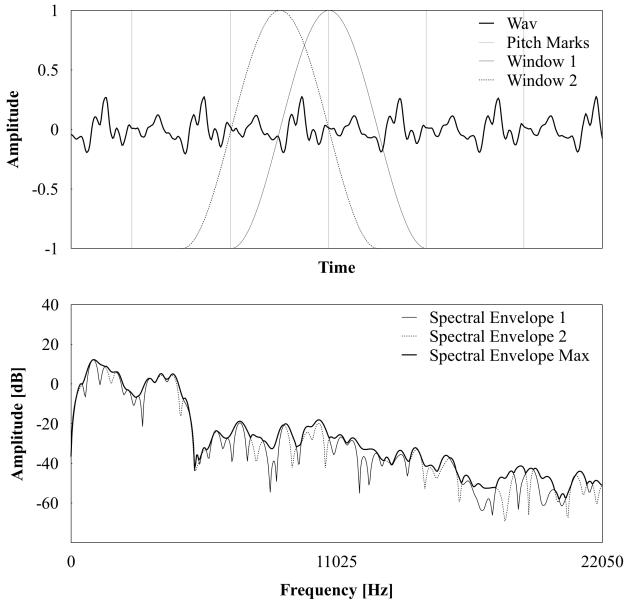

図 4 2 つの単位波形の抽出（上），2 つの振幅スペクトルと，それらの最大値の算出 （下）

分析ピッチマークの配置位置は，分析 F0 より算出される声帯振動の生じる時刻 [7] に基づき決定する．一般的にピッチマーク分析では，波形のローカルピークに合わせてピッチマーク位置を決定するなど，何らかの分析処理を伴うことが多い．一方，本手法では，分析 F0 のみからピッチマーク配置位置を決定する，シンプルな方法を採用している．

2.2.3 スペクトル包絡推定

スペクトル包絡は，分析ピッチマークごとに推定される．まず，対象となる分析ピッチマークを中心とし，その前後に隣接する分析ピッチマークまでの，左右非対称のハニング窓を用いて単位波形を抽出する．同様に，分析窓の中心を窓幅の 1/4 サイズ分ずらした位置からも単位波形を抽出する（**図 4** 上）．次に，抽出された 2 つの単位波形より，FFT 点数 N で，それぞれの振幅スペクトルを算出し，2 つの振幅スペクトルにおける周波数ビンごとの最大値を取ったものを，推定スペクトル包絡とする（図 4 下）．

単一の分析窓から得られたスペクトル包絡には谷が存在する [5] ため，本手法では，2 つの単位波形を用いることで，スペクトル包絡の谷を効率よく埋めている．

2.2.4 非周期性指標推定

非周期性指標は，スペクトル包絡と同様に，分析ピッチマークごとに推定される．また，本手法における非周期性指標は，スペクトル形状で扱われる．

まず，対象となる分析ピッチマークを中心とする，窓幅 W，FFT 点数 N の STFT に基づき振幅スペクトルを求める．次に，2.2.1 の (a) で，包絡成分と微細構造を分離し，更に，2.2.1 の (c) を用いて，微細構造を再度分離することで，対数振幅上において，山谷の中心が概ね 0 dB 付近を辿るように補正された微細構造を獲得する（**図 5** の実線）．

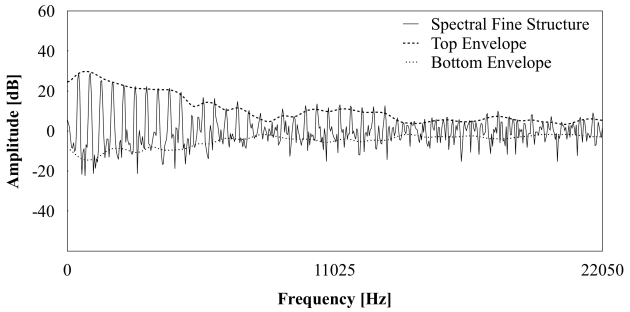

図 5 微細構造の上側包絡と下側包絡

この補正された微細構造の上側包絡と下側包絡を求め(図 5 の破線と点線)，対数振幅上で下側包絡から上側包絡を減じた上で対数を外し，推定非周期性指標とする．図 5 において，上側包絡と下側包絡の差が少ないほど，その帯域の非周期性指標が高いことを意味する．

上側包絡は，微細構造の 0 dB 以上の成分（線形振幅上で，1.0 以上の成分）のみを取り出し，2.2.1 の (b) を利用して分離するという処理を，再帰的に行うことで獲得される．これは，改良ケプストラム法 [13] におけるケプストラムを，パワースペクトルの逆フーリエ変換に置き換えた手法と位置付けられる．

下側包絡は，微細構造の 0 dB 以下の成分（線形振幅上で，1.0 以下の成分）のみを取り出し，対数上の正負を反転させた上で上側包絡と同様に計算し，対数上の正負の反転を戻す．ただし，下側包絡を求める際，微細構造の谷が極端に小さな値をとることがあるため，対数上の正負を反転させた際に，事前に求められた上側包絡を参照し，上側包絡を一定以上超えた値を圧縮することで，微細構造の谷の影響を抑制している．

2.3 合成アルゴリズム

2.3.1 合成 F0 と合成ピッチマーク作成

合成用 F0 は，分析 F0 と F0 エフェクト処理に応じて生成される．F0 エフェクト処理を何も行わない場合は，分析 F0 がそのまま合成 F0 となる．

合成ピッチマークの配置位置は，合成 F0 より算出される声帯振動の生じる時刻に基づき決定する．また，合成ピッチマークは，参照する分析ピッチマークを必ず 1 つもっており，その合成ピッチマーク以前の時刻で最も近い分析ピッチマークを参照する．

2.3.2 周期成分・非周期成分の算出

分析ピッチマーク位置で推定されたスペクトル包絡と非周期性指標に対し，エフェクト処理を行った上で，周期成分・非周期成分を算出する．スペクトル包絡を $H(\omega)$，非周期性指標を $A_p(\omega)$ とすると，周期成分は $|H(\omega)|(1-A_p(\omega))$，非周期成分は $|H(\omega)|A_p(\omega)$ で算出される [7]．続いて，周期成分と非周期成分にエフェクト処理が行われる．

2.3.3 最小位相応答による波形合成

合成ピッチマークごとに，参照する分析ピッチマークから得られる周期成分・非周期成分を用いて，単位波形を生成し重畳加算する．その際，周期成分はパルス，非周期成分はベルベットノイズ [14], [15] を励起信号とし，最小位相応答を畳み込むことで波形を生成する．

非周期成分の合成には，ホワイトノイズを畳み込むのが一般的 [7] だが，ベルベットノイズは同等の品質で合成することが可能な上，ホワイトノイズよりも操作性が高いことから，本手法では，ベルベットノイズを採用している．

なお，CHERRY PIE では，設定でホワイトノイズに切り替えることも可能である．今後は，ベルベットノイズの操作性を活かしたエフェクトも検討したいと考えている．

2.4 リアルタイム音声分析合成を実現するための観点

高度な解析を必要とする音声分析合成において，低遅延のリアルタイム処理を実現するためには，分析精度の追求よりも効率を重視した考え方が重要となる．また，基本的に，分析の誤りを補正することを前提にはできないため，入力音声に対して頑健性の高い分析方法，及び，合成方法を確立する必要がある．それらを実現するための本手法における方針をいくつか述べる．

2.4.1 F0 推定における方針

本手法では，アルゴリズム遅延を低くするために，単一フレームの短時間波形のみで利用可能な F0 推定アルゴリズムを採用している．ただし，現在の推定 F0 候補の選択の際に，過去フレームの推定 F0 値を参照し，候補の優先度に反映するなど，アルゴリズム遅延に影響を与えない形での精度向上を図っている．

F0 推定では，半ピッチ，倍ピッチ誤りのような，大きな推定誤りを少なくすることを重視し，合成品質に与える影響が比較的小さいと考えられる微細な推定誤差については重視していない．つまり，微細な推定誤差を小さくするよりも，全体的な安定性の維持を優先している．本手法の F0 推定アルゴリズムは，単一フレームの短時間波形のみで利用できる上，ノイズに対する頑健性が高く，動作が安定している．

2.4.2 有声・無声区間判定における方針

本手法では，F0 推定と同時に有声・無声区間判定を行なっているが，合成時には有声・無声情報を用いていない．精度の高い有声・無声区間判定を行うには，未来の情報と合わせた補正処理が必要となり，その場合はアルゴリズム遅延が増大する．また，有声区間を無声区間と誤判定した場合は，合成音声の品質が著しく劣化する．更に，有声区間と無声区間の隣接する箇所で，非連続性からノイズが生じる可能性がある．これらの理由より，本手法では頑健性の維持を優先して，有声・無声情報を用いずに，全てを有声区間として処理を行う．

本来の無声区間は，非周期成分が支配的になることから非周期性指標が高く推定され，有声区間として処理を行っても品質劣化は比較的少ないと考えられる．また，ヒューリスティックな手法ではあるが，スペクトル重心の値に応じて非周期性指標を増幅させるなどの処理を行い，無声音の合成品質の向上を図っている．現在は，CHERRY PIE の GUI 上の F0 グラフ表示にのみ，有声・無声情報が反映されている．

VOCODER 方式の音声分析合成では，例えば，Janis Joplin のような声を歪ませた歌唱法において，半ピッチ誤りや，有声区間を無声区間とする誤判定により，品質劣化が起きやすい傾向にあるが，本手法は品質の安定性が高い感触が得られている．

2.4.3 スペクトル包絡・非周期性指標推定における方針

本手法では，分析ピッチマークごとにスペクトル包絡・非周期性指標推定を行っており，分析ピッチマーク間における，スペクトル包絡・非周期性指標の補間は行っていない．分析ピッチマーク間の補間を行うためには，1 つ先のスペクトル包絡・非周期性指標が必要となり，アルゴリズム遅延が増大する．また，分析ピッチマーク間の補間の有無が品質に与える影響は比較的小さいと考えられる．これらの理由より，本手法では効率を重視し，合成ピッチマークが 2 つの分析ピッチマーク間に位置する場合は，その合成ピッチマーク以前の時刻で，最も近い分析ピッチマークの位置で推定されたスペクトル包絡・非周期性指標を利用する．

3. リアルタイム音声分析合成エフェクター CHERRY PIE

VST 及び，AU で実装された，CHERRY PIE の GUI を**図 6** に示す．GUI 上部にはグラフ画面が搭載されており，合成された単位波形，F0，スペクトル包絡，非周期性指標，周期成分，非周期成分のグラフを表示することができる．また，グラフ画面の右に位置する縦スライダーでは，F0 の推定範囲を設定することが可能である．F0 のグラフ表示は F0 推定範囲スライダーと連動しているため，設定する上で参考になる．

グラフ画面左下に位置する「MASTER」項目内では，エフェクトレベル，F0 推定精度設定，ファクトリープリセットが設定できる．エフェクトレベルは 0.0〜1.0 で設定し，0.0 にすると，加工されていない分析合成音声が出力される．F0 推定精度設定は，処理負荷と推定精度に応じて 5 段階で選択できる．ファクトリープリセットには，予め準備されたエフェクトセッティングが実装されている．

「MASTER」項目以降は，CHERRY PIE に搭載された各エフェクトを設定する項目になっている．

図 6 CHERRY PIE の GUI

3.1 CHERRY PIE 搭載エフェクト

3.1.1 F0 エフェクト

F0 エフェクトは，「PITCH」項目内より設定可能となっている．ここでは，相対的な F0 変化を与えるエフェクト（ピッチシフト）や，指定した F0 で合成するエフェクトが実装されている．また，GUI 下部のキーボードのキーを押している間は，キーに応じた F0 で合成され，楽器のように操作することができる．音声分析合成が有する高い柔軟性により，品質の高い F0 操作が可能である．

3.1.2 スペクトル包絡・非周期性指標エフェクト

スペクトル包絡・非周期性指標エフェクトは，「SPECTRAL ENV / APERIODICITY」項目内より設定可能となっている．ここでは，スペクトル包絡・非周期性指標に対し，3 バンドの帯域を設定し，各バンドごとに値を増減させるフィルタが実装されている．その際，スペクトル包絡は，各バンドごとに dB 単位で値を設定する．

非周期性指標は，各バンドごとに-1.0〜1.0 までの値を設定し，0.0 の時は変化せず，-1.0 の時は-60 dB，1.0 の時は 0 dB となるように設計されている．なお，-1.0〜0.0 の間，及び，0.0〜1.0 の間は，対数上で線形に変化する．

更に，スペクトル包絡・非周期性指標に対する，周波数軸方向の伸縮エフェクトも搭載されている．これは，一般的なフォルマントシフターに相当するものだが，CHERRY PIE では，高域は原形を保ったまま，低域・中域を効果的に操作するという観点から，区分線形伸縮を用いている．この伸縮処理は，対数振幅上で行われる．

3.1.3 周期成分・非周期成分エフェクト

周期成分・非周期成分エフェクトは，「PERIODIC ENV / APERIODIC ENV」項目内より設定可能となっている．ここでは，周期成分・非周期成分に対し，3 バンドの帯域を設定し，各バンドごとに dB 単位で値を増減させるフィルタが実装されている．更に，周期成分・非周期成分のそれぞれの振幅を，全て 0.0 にする機能が実装されている．この機能を利用することで，周期成分のみ，及び，非周期成分のみの合成音声を簡単に確認することができる．

前述したスペクトル包絡・非周期性指標エフェクトでは，周期成分・非周期成分の両方に影響を与えるが，周期成分・非周期成分エフェクトでは，それぞれ独立した変形を行うことが可能である．

3.1.4 DNN 声質変換エフェクト

DNN (Deep Neural Networks) 声質変換エフェクトは，「NN VOICE CONVERTER」項目内より設定可能となっている．DNN 声質変換エフェクトでは，予め，音声データ（パラレルデータ）を深層学習することで得られた学習結果（ネットワークファイル）を，CHERRY PIE 上に読み込み，入力音声のスペクトル包絡を変形して，別人のような音声に変換することができる．

ネットワークファイルには，声質変換モデルとなるニューラルネットワークと，変換アルゴリズムに関する情報が記載されており，CHERRY PIE の GUI 上にドラッグ＆ドロップ，もしくは，読み込みボタンを利用して読み込む．なお，ネットワークファイルは 2 つまで読み込める．

DNN 声質変換エフェクトには，スペクトルモーフィングの考え方を取り入れており，入力音声のスペクトル包絡と変換音声のスペクトル包絡に対し，3 バンドの帯域を設定し，各バンドごとに比率を変えてモーフィングすることができる．これは，操作性の向上に加え，出力音声の品質の向上にも有効である．

DNN 声質変換エフェクト内の「MATCHING」スライダーを操作すると，入力音声のスペクトル包絡に区分線形伸縮（3.1.2 と同様の処理）が施され，簡易的に多対一声質変換に拡張することが可能となる．CHERRY PIE のデモムービー [10] 内の，DNN 声質変換エフェクトで利用されているネットワークファイルは，「女性声優音声 → 女性バーチャルシンガー音声」のパラレルデータで学習されており，それを，学習には利用されていない，全く別人の英語圏の女性歌手・男性歌手に適用することで，多対一声質変換を実現している．

声質変換モデルとなるニューラルネットワークの学習とネットワークファイルの生成は，別途，学習機能を搭載したアプリケーションを用いて行う（**図 7**）．本アプリケーションでは，2 つのオーディオに対する，DTW (Dynamic Time Warping) を算出するためのインターフェースが実装されている．DTW の正確性を重視したい場合は，GUI を利用して対応範囲を指定することが可能である．学習後に，入力，目標，変換結果のスペクトル包絡を比較するためのグラフ機能も搭載されている．

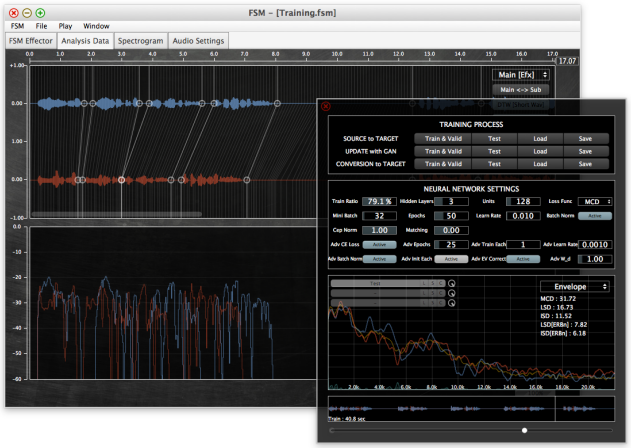

図 7 学習機能を搭載したアプリケーション

声質変換手法と，それに伴うニューラルネットワークの構成については，スペクトル特徴量変換，差分スペクトル補正 [16], [17] などの考え方を基に，独自の工夫を加えた複数の変換アルゴリズムが実験的に実装されている．変換アルゴリズムの選択や設定は，ニューラルネットワークのハイパーパラメータと共に GUI 上から行う．変換アルゴリズムに関する情報は，学習結果として生成されるネットワークファイル内に記載されており，DNN 声質変換エフェクト側では，利用するネットワークファイル内の情報に応じて，変換アルゴリズムと各種パラメータを切り替える．

これらの変換アルゴリズムでは，フルバンドでの声質変換を効率的に実現するため，低域から中域までのスペクトル包絡から求めたメルケプストラムと，全域のスペクトル包絡から求めたケプストラムを算出し，それぞれ別々のニューラルネットワークで学習している．DNN 声質変換エフェクト側では，低域から中域までのスペクトル包絡の変換結果と，全域のスペクトル包絡の変換結果を，周波数上でクロスフェードさせて混合する．こうして，サンプリング周波数が 44,100 Hz 以上の音声においても，ダウンサンプリングすることなく，効率的に声質変換を行うことが可能である．

更に，本アプリケーションでは，学習済みのネットワークファイルに対して敵対的学習を適用し，声質変換モデルをアップデートする機能を搭載している．敵対的学習のアーキテクチャは，Saito[18] らの手法を参考にした．

近年は，リアルタイム声質変換の研究 [19], [20], [21] が盛んになり，既に市販されているアプリケーション [22] も存在する．その中でも，CHERRY PIE は，遅延の低い部類に入ると考えられる．更に，CHERRY PIE の声質変換は，モーフィングやマッチング機能を取り入れることで，他手法より操作性が高いものとなっている．

4. おわりに

本稿では，音声分析合成エフェクター CHERRY PIE を基に，リアルタイム音声分析合成を実現するためのアルゴリズムの骨格，及び，考え方について述べた．

今後は，現状のリアルタイム性の枠組みを維持しつつ，品質の向上に取り組む予定である．また，VOCODER 方式の音声分析合成特有の効果性を，より発揮するエフェクターに発展させていきたいと考えている．

現在は，CHERRY PIE 上に，F0 変化とスペクトル包絡変化が連動するビブラートや，VOCAL DRIVE[9], [10] と類似した効果の音声を歪ませるエフェクトを実験的に実装している．更に，複数の音声を入力し，F0 やスペクトル包絡・非周期性指標をリアルタイムでモーフィングするエフェクトも検討中である．また，アルゴリズム遅延をより低減する方式に向けた検討を，更に進める予定である．

本手法の有する実用性と頑健性の高さは，録音済みの音声編集に特化したアプリケーションにおいても有効であると考えられる．予め録音された音声を対象とする場合，発声タイミングや音長など，時間軸方向に対する編集が可能となるため，より高い操作性が得られる．VOCODER 方式の高い柔軟性を活かした，高度な編集を行うことができる音声編集アプリケーションの開発の検討も行いたい．

謝辞 本研究を遂行するにあたり，貴重なアドバイスを頂いた，産業技術総合研究所 後藤真孝氏，中野倫靖氏に深謝する．

本研究の遂行をサポートして頂いた，クリプトン・フューチャー・メディア株式会社 佐々木渉氏，黒田毅氏，石岡卓也氏，山根壮一氏に深謝する．

参考文献

[1] H. Kawahara and M. Morise, “Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework,” SADHANA - Academy Proceedings in Engineering Sciences, vol. 36, no. 5 , pp. 713-728, Oct. 2011.

[2] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno, “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation,” Proc. ICASSP 2008, pp. 3933-3936, Las Vegas, March 30 - April 4, 2008.

[3] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016.

[4] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol. 84, pp. 57-65, Nov. 2016.

[5] Tomoyasu Nakano, Masataka Goto, “A Spectral Envelope Estimation Method Based on F0-Adaptive Multi-Frame Integration Analysis,” Proc. SAPA-SCALE 2012, pp.11-16 September 2012.

[6] Tomoyasu Nakano, Masataka Goto, “VocaRefiner: An Interactive Singing Recording System with Integration of Multiple Singing Recordings,” Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013), pp.115-122 August 2013.

[7] 森勢将雅（日本音響学会編），“音声分析合成，” コロナ社, ISBN 978-4-339-01137-1, July 2018.

[8] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, “Implementation of realtime STRAIGHT speech manipulation system,” Acoust. Sci. & Tech., vol. 28, no. 3, pp. 140-146, March 2007.

[9] Labopton BLOG
https://blog.crypton.co.jp/l/2019/03/vocefx/

[10] Vocal Drive & Cherry Pie Demo Movie
https://www.youtube.com/watch?v=YDYMPZXiM60

[11] 嵯峨山茂樹, 古井貞熙, “ラグ窓を用いたピッチ抽出の一方法，” 電子情報通信学会全国大会予稿集, 1235, Vol. 5, p. 263, 1978.

[12] Yukara Ikemiya ,Katsutoshi Itoyama, Kazuyoshi Yoshii, “Singing Voice Separation and Vocal F0 Estimation Based on Mutual Combination of Robust Principal Component Analysis and Subharmonic Summation,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), Volume 24 Issue 11, pp. 2084 - 2095, November 2016.

[13] 今井聖，阿部芳春，“改良ケプストラム法によるスペクトル包絡の抽出，” 電子通信学会論文誌，Vol. J62-A, No. 4, pp. 217-223, 1979.

[14] Matti Karjalainen and Hanna Järveläinen, “Reverberation modeling using velvet noise,” AES 30th International Conference, Saariselkä, Finland, 2007 March.

[15] Hideki Kawahara, Ken-Ichi Sakakibara, Masanori morise, Hideki Banno, Tomoki Toda, Toshio Irino, “Frequency domain variants of velvet noise and their application to speech processing and synthesis,” Proc. Interspeech 2018.

[16] K. Kobayashi, T. Toda, S. Nakamura, “F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential,” Proc. IEEE SLT, pp. 693-700, Dec. 2016.

[17] T. Toda, A.W. Black, K. Tokuda. “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp. 2222-2235, Nov. 2007.

[18] Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari, “Evaluation of DNN-Based Voice Conversion Deceiving Anti-spoofing Verification,” IEICE Technical Report, SP2016-69, vol. 116, no. 414, Jan., 2017.

[19] T. Toda, T. Muramatsu, H. Banno. “Implementation of computationally efficient real-time voice conversion,” Proc. INTERSPEECH, Portland, USA, Sep. 2012.

[20] 荒川陸，高道慎之介，猿渡洋，“リアルタイム DNN 音声変換の実装とデータ拡張法による音質改善法，” 日本音響学会 2019 年春季研究発表会講演論文集, 1-10-4, 2019.

[21] 廣芝和之，能勢隆，宮本颯，伊藤彰則，小田桐優理，“畳込みニューラルネットワークを用いた音響特徴量変換とスペクトログラム高精細化による声質変換，” 音学シンポジウム 2018.

[22] Voidol
https://crimsontech.jp/apps/voidol/