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
SP2016-62 (2016-12)

高い雑音耐性と推定精度を両立する基本周波数推定法の提案と評価

森勢　将雅†

† 山梨大学大学院総合研究部　〒 400–8511 山梨県甲府市武田 4–3–11
E-mail: †mmorise@yamanashi.ac.jp

あらまし　基本周波数 (F0，最近は FO と表記することもあるが本稿では F0 に統一する) は，周期的に生じる声帯振動間隔の最も短いものの逆数として定義され，知覚する音声の高さに概ね対応する音声の主要なパラメータである．F0 は様々な音声処理に利用されるパラメータであり，例えば Channel vocoder の考えに基づいた高品質音声合成では，音声から F0 を可能な限り高い精度で推定することが要求される．筆者らは，これまで高 SNR の音声を対象とした実時間処理が可能な推定法について検討し，SNR が 30 dB 以上であれば実時間処理が可能であり，かつ最新の方法と比較しても遜色ない性能が達成可能な方法を提案してきた．一方，例えば統計的音声合成では，学習に必要な音声パラメータは事前に分析しておけば良いため，実時間性よりも高い精度と雑音に対する頑健性を備えた方法が望ましいといえる．本稿では，計算速度ではなく，高い耐雑音性と推定精度にフォーカスを絞った F0 推定法 Harvest を提案する．Harvest は，音声スペクトルが調波構造を持つことに着目し，基本波に相当するピークを検出する方法を採用している．まず，高調波と低域雑音を除去するため，様々な中心周波数のバンドパスフィルタによるフィルタリングを実施し，得られた多チャネル信号から F0 の可能性がある候補を全て選定する．その後，選定された候補を瞬時周波数を用いて補正し，時系列の連続性を考えて接続することで最終的な F0 軌跡を生成する．本稿では，音声データベースを用いた評価，および筆者らが 2016 年に提案した耐雑音性評価法により提案法の有効性を示す．
キーワード　音声分析，基本周波数，基本波，耐雑音性

Proposal of a robust and high-performance F0 estimator and its evaluation

Masanori MORISE†

† Interdisciplinary Graduate School, University of Yamanashi,　4-3-11, Takeda, Kofu, 400–8511, Japan
E-mail: †mmorise@yamanashi.ac.jp

Abstract　Fundamental frequency (F0) is related with the perceived pitch of the periodic signal and is one of the most important parameters for various kinds of speech processing. Modern channel vocoders for high-quality speech synthesis generally require high-performance estimators in speech parameters including F0. We have proposed a rapid and reliable F0 estimator for real-time applications. On the other hand, other applications such as statistical parametric speech synthesis require a robust estimator rather than computational cost. This paper presents a robust and high-performance F0 estimator named *Harvest* for high-quality speech synthesis. The proposed estimator consists of three steps: multi-channel band-pass filtering with different center frequencies, calculation of F0 candidates and connection on the basis of the continuity of the F0 contour. Two evaluations in estimation performance and noise robustness were carried out compared with other modern estimators. The result showed that the proposed estimator was superior to others in both estimation performance and noise robustness.
Key words　Speech analysis, fundamental frequency, fundamental component, noise robustness

1. ま え が き

基本周波数 (F0) は人間が周期信号から知覚する高さに概ね対応するパラメータであり，音声波形からの F0 推定は長期にわたり取り組まれてきた研究分野である．例えば，Channel vocoder [1] の考え方に基づく音声分析合成技術では，イントネーションの加工を可能にする音声変換が可能になる．また，F0 は，ピッチ同期分析 [2] というスペクトル解析法にも利用され，音声を構成する他のパラメータであるスペクトル包絡の推定や [3]〜[5] や非周期性指標推定 [6] にも利用されることから，

高精度な F0 推定法は他の音声パラメータ推定の精度改善にも貢献する．さらに，音声に限定されず周期信号であれば F0 が存在するため，F0 推定法は，周期性を有する楽器音の解析にも利用することが可能である [7]．

筆者らは，高品質音声分析合成法に関する検討を進めており，高品質音声分析合成では主に防音室等で録音された音声が利用されることから，高 SNR の音声を対象とした実時間処理向けの方法を提案してきた [8], [9]．一方，統計的音声合成 [10] では，音声の分析は合成と独立にオフラインで行えばよいことから，実時間性よりも雑音を含む音声でも頑健に動作する方法の提案が望まれていた．現在筆者が開発を進めている WORLD [11] (D4C edition [6]) は，The Merlin toolkit [12] など統計的音声合成用の音声分析部にも利用されているが，耐雑音性に優れた方法ではないため，雑音を含む音声での品質劣化が課題であった．

本稿では，WORLD で分析可能な音声の範囲をさらに広めるため，現在用いている F0 推定法に加え，主に低域の雑音に頑健な方法 Harvest を提案する．音声には，録音環境に起因する雑音だけではなく，声がかすれているなどの要因で雑音が混入するため，耐雑音性の向上は，WORLD の品質をさらに高める効果も期待できる．以下では，F0 推定の関連研究と提案法との位置づけを明確にし，提案法である Harvest の具体的な処理を説明する．音声データベースを用いた評価，および人工音を用いた耐雑音性の評価から，提案法の有効性について議論する．

2. F0 推定法の関連研究

音声波形からの F0 推定法には多数の方法が提案されており，古くは波形の相関を用いた方法 [13] と，ケプストラム [14], [15] に代表されるパワースペクトルの特徴に着目した方法に大別される．波形の相関を用いた方法はその後 YIN [16] や pYIN [17] の提案に繋がり，高い性能を達成可能となった．パワースペクトルの特徴に着目した方法では，2008 年に SWIPE が高精度な方法として提案されている [18]．

その他の特徴量に関しても，ウェーブレット変換を用いた方法 [19] や，声帯振動の生じる時刻を直接計算して F0 推定に利用する方法 [20]，瞬時周波数 [21] に着目した方法などが提案されてきた．耐雑音性に特化した方法も検討がなされており，いくつかの方法が提案されている [22], [23]．WORLD で採用している DIO [8], [9] は，周期信号のスペクトルが調波構造となることに着目し，そのもっとも低い周波数ピークを基本波として抽出する基本波検出法 [24] をベースとしている．基本波のみを残す低域通過フィルタは F0 が未知である以上設計できないため，低域通過フィルタのカットオフ周波数を様々な周波数に設定し，フィルタリング後の波形から特徴量を求め，信頼できる F0 候補を選定する手順を採用している．現在，さらに速度を追求した DIO の改良法が大道らにより提案されており [25]，低域通過フィルタ後に F0 候補を選定するための特徴量についても研究された事例がある [26]．

本稿で提案する Harvest は，基本波検出法である DIO を基準に改良を加えた方法である．DIO では低域通過フィルタを用いていたため，基本波より低い低域雑音に弱いことが課題であった．Harvest は，低域通過フィルタの代わりに，文献 [27] に記載されている方法と同様の帯域通過フィルタ群を用いることで低域雑音が結果に与える影響を低減する．また，フィルタリングにより得られた全 F0 候補を「ぼかす」処理，および周波数軸で定義される別の特徴量により修正し，さらに F0 軌跡の連続性に着目した接続を行うことにより，耐雑音性の向上を実現する．以下の章では，Harvest による F0 推定法の流れについて示す．

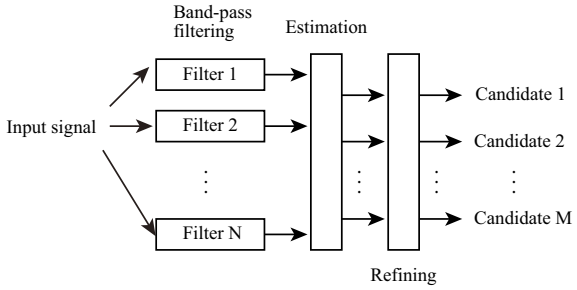

図 1　Harvest による F0 候補選定の流れ．推定された M 個の候補から適切な F0 軌跡を生成する．

3. 提案法 Harvest の概要

Harvest は F0 候補の推定部と推定された F0 候補から軌跡を生成する部から構成される．また，F0 軌跡は 1 ms 毎に計算する必要がある．図 1 は，Harvest による F0 候補推定の流れを示す．F0 候補の推定は，DIO と同様にフィルタリングによる基本波検出法をベースにしているが，DIO では低域通過フィルタを利用する一方，Harvest では帯域通過フィルタを利用する．以下ではいくつかのマジックナンバーを用いて説明するが，これらは，次章で実施する音声データベースを用いた評価で最小の誤差となるようチューニングした結果である．

3.1 F0 候補の推定

3.1.1 帯域通過フィルタによる基本波検出

Harvest では，第一ステップとして，中心周波数が ω_c Hz の帯域通過フィルタを窓関数と cos 波を組み合わせにより設計し，基本波の抽出を試みる．帯域通過フィルタの設計については，文献 [27] の方法と同様である．

$$h(t) = w(t)\cos(\omega_c t), \tag{1}$$

ここで，$w(t)$ は Nuttall 窓 [28] を表し，窓長は $4T_0 (T_0 = 2\pi/\omega_c)$ である．窓が存在する時間区間は $-2T_0$ から $2T_0$ であり，時刻 0 で振幅がピークとなる特徴を有する．この窓長により設計された Nuttall 窓は，スペクトルの最初のゼロ点が ω_c Hz で生じる．このような窓関数に周波数が ω_c Hz の cos 波を乗算することで，図 2 に示すように，中心周波数が ω_c Hz で $0, 2\omega_c$ Hz にゼロ点を持つ帯域通過フィルタが設計できる．

Harvest では，F0 の探索範囲を探索範囲の下限から上限まで 40 ch/octave に中心周波数を設定した帯域通過フィルタ群によるフィルタリングを行う．得られた多チャネル信号に対し，DIO と同様に 4 つの周期 (正から負，負から正へのゼロクロス

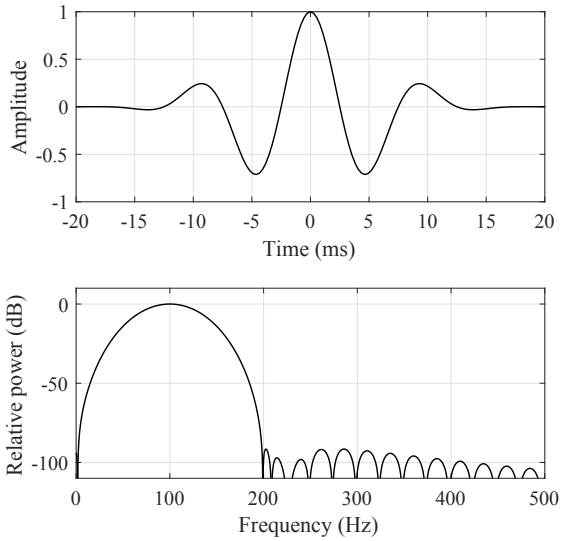

図 2　基本波検出に用いる帯域通過フィルタの例．この図では，ω_c を 100 Hz に設定している．

周期，および波形のピークと谷の周期) を各時刻について算出する．基本波のみが抽出された場合，フィルタ後の波形は sin 波となるため，4 つの周期は等しい値を示す．これらの手続きは文献 [8], [9] に示されている方法と同様である．

Harvest は，DIO と同様に 4 つの周期の平均値の逆数を全信号・各時刻について計算することになるが，計算された結果が $\omega_c \pm 10\%$に入らない場合 0 と置き換える処理を実施する．この処理は，音声のパワースペクトルピークが帯域通過フィルタの中心周波数近辺に存在する場合にのみ F0 候補が検出されることを示す．各時刻について最大でバンドパスフィルタの個数分の候補が検出されるため，ここから F0 候補の選定を実施する．

3.1.2　F0 候補の選定

図 3 は，横軸を帯域通過フィルタの中心周波数，縦軸を得られた F0 候補（4 つの周期の平均の逆数）として表示したグラフを示す．前節で述べたとおり，中心周波数と F0 候補が合致する周波数のみ値を有する．ただし，分析信号が調波構造を有する場合，中心周波数 ω_c が F0 より 10%以上離れていても F0 に起因するピークは顕著である．つまり，ω_c が F0±10%の範囲では F0 の値を示し，±10%以上の範囲から次の調波が支配的になる周波数までは 0 を示すこととなる．Harvest では，この特性に着目し，横軸について一定区間値を有する帯域を検出し，その帯域における値の平均値を F0 候補とする．図 3 の場合，推定すべき F0 は約 145.4 Hz で選出されているが，その 2 倍，3 倍の周波数を含めたいくつかの候補が検出されている．

ここまでで得られた全 F0 候補について，次のステップでは瞬時周波数による補正を行い各候補の信頼度を算出する．ただし，雑音が局所的に加わり特定フレームでの推定ができなかった可能性を勘案し，n フレームで得られた F0 候補を n±3 フレームにコピーする処理を実施する．特定のフレームを前後のフレームにコピーする処理は，F0 候補を時間的に「ぼかす」処理といえる．

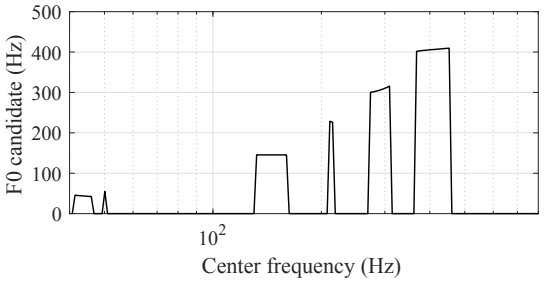

図 3　横軸を帯域通過フィルタの中心周波数，縦軸を得られた F0 候補として表示したグラフ．スペクトルにおける顕著なピークのみ検出可能である．

瞬時周波数（角周波数表現）は，以下に示す Flanagan の式 [29] により計算する．

$$\omega_i(\omega,t) = \frac{\Re[S(\omega,t)]\Im[\frac{\partial S(\omega,t)}{\partial t}] - \Im[S(\omega,t)]\Re[\frac{\partial S(\omega,t)}{\partial t}]}{|S(\omega,t)|^2}, \tag{2}$$

ここで，$S(\omega,t)$ は，時刻 t で窓関数により切り出された入力波形のスペクトルに対応する．波形の切り出しは，F0 候補の逆数の 3 倍を窓長とする Blackman 窓により行う．切り出された波形から瞬時周波数とパワースペクトルを計算し，以下の式により修正された F0 候補を計算する．

$$\hat{\omega}_0(t) = \frac{\sum_{k=1}^{K} |S(k\omega_0,t)|\omega_i(k\omega_0,t)}{\sum_{k=1}^{K} k|S(k\omega_0,t)|}, \tag{3}$$

ここで，$\omega_0(t)$ と $\hat{\omega}_0(t)$ は，それぞれ修正前の F0 と修正された F0 の角周波数表現を表す．瞬時周波数の性質より，$k\omega_0$ Hz における瞬時周波数は $k\omega_0$ Hz となる．各調波の周波数に対応する瞬時周波数から $k\omega_0$ を引けば，理論的には全ての調波の周波数で 0 を示すことになる．本稿では，K 個の調波について上記のパラメータの 2 乗を計算し，結果から平均の逆数を求め各 F0 候補の信頼度とした．ただし，音声の高域は非調波的でありパワーも小さいことから，低域の調波数個に限定する必要があるため，本稿では，K=6 に設定した．

3.2　F0 軌跡の生成と平滑化

ここまでの処理により，各フレームにおける多数の F0 候補と各候補の信頼度が算出される．まずは全フレームについて最も信頼度が高い F0 で構成した軌跡を基準として，滑らかな F0 軌跡となるよう繋げる操作を行う．この操作は，文献 [30] を基準にしているが，詳細な処理とパラメータが異なる．

（1）　1 フレーム前の F0 候補との差が 1%以上ある場合，その候補を 0 に修正

（2）　0 以外の値を持つ候補が連続して 5 ms 分存在しない場合，その区間の候補全てを 0 に修正

（3）　有声音から無声音に切り替わる境界における F0 を基準にし，有声区間に隣接する無声区間で基準 F0 の ±a%以内に候補が存在する場合，その中で最も基準 F0 に近いものを候補

として採用

(3) は，有声区間から無声区間へ切り替わる境界の F0 候補を基準に，隣接する無声区間の F0 候補から適切な候補を選び有声区間を延長する操作といえる．$\pm a$%に候補が無い場合，$\pm$14%に範囲を広げて探索し，該当する候補が 1 つの場合は採用する．a は 7%が基準値であるが，あるフレームで候補が検出された場合，その次の探索では，7%に基準 F0 と検出された候補の変化率を加算することで動的に変化させる．$n+1$ ms に適切な候補がない場合，$n+3$ ms を上限に探索を続け，$n+3$ ms まで候補が検出されない，あるいは有声音区間の延長が 100 ms に達するまで再帰的に繰り返す．

この接続は有声音区間単位で行うため，一部の無声音区間が前後の有声音区間の延長により重複して F0 を持つすることとなる．この際は，重複区間における全 F0 候補の信頼度の総和を求め，より大きい区間の全候補を採用する．その後，9 ms 以下の無声音区間については，境界部分の F0 を用いて線形補間により有声区間へと修正する．例えば n と $n+5$ ms に F0 が存在し間が無声区間の場合，$n+1$ から $n+4$ ms の F0 は n ms と $n+5$ ms の F0 の線形補間により値を与える．

こうして得られた F0 軌跡を低域通過フィルタにより処理した結果が最終的な F0 軌跡となる．ここではカットオフを 30 Hz に設定した 2 次のバタワースフィルタにより F0 軌跡を処理する．処理した結果を時間反転し再度同一のフィルタで処理し，処理後の信号を再度時間反転することで，位相特性がフラットな低域通過フィルタを実現することとした．

4. 評　　価

F0 推定法の評価は，一般的に実音声と Electroglottograph (EGG) 信号を同時収録した音声データベースにより行われる．EGG 信号は，喉に電極を張り付けて声帯の開閉に対応して変化する値を観測した時系列信号のため，声道形状による影響を含まない特性を有する．本稿では，さらに，人工的に生成した F0 が既知の信号を用いた評価法 TUSK [31] を用いて Harvest を比較評価する．

4.1 比較する F0 推定法

比較する手法には，いくつかの高精度な F0 推定法を用いることとした．時間軸の特徴から F0 を推定する YIN [16]，およびスペクトルの特徴から推定する SWIPE [18] は，各論文の著者が実装した MATLAB プログラムが存在するためそれらを利用した．高品質音声合成に利用されている方法では，STRAIGHT [32] の最新版で採用している NDF [33] と，TANDEM-STRAIGHT [34], [35] で採用している XSX を用いた．また，Harvest の前身といえる DIO [8], [9] も評価に加えることとした．

4.2 音声データベースによる評価

音声データベースによる評価では，阿竹らにより構築された男女 7 名，計 14 名から構成される音声計 840 発話からなるデータベース [36] を用いた．本データベースに収録された音声のサンプリング周波数は 16 kHz であり，EGG 信号と有声・無声判定のラベル情報が同時に収録されている．EGG 信号は波形であり F0 軌跡そのものではないため，本稿では，EGG 信号の時間差分信号を対象に NDF で全フレームの F0 を推定し，データベースに収録された有声区間を対象として推定された F0 を真値とした．評価指標には，Gross error [16] や Fine pitch error [37] が提案されており，本評価では Gross error を用いることとした．全 F0 推定法で F0 推定範囲の下限を 40 Hz，上限を 800 Hz に設定し，シフト幅は 1 ms に設定した．

実験結果は，Gross error の小さい順に Harvest が 0.39%，NDF が 0.91%，XSX が 0.97%，SWIPE が 1.19%，DIO が 1.33%，YIN が 1.84%であった．本結果は，Harvest の性能が他の方法よりも優れていることを示す．一方，EGG 信号からの推定値やデータベースの有声無声判定に誤差が含まれる可能性もあり，それらを含めた有効性の考察は考察の節で述べる．

4.3 耐雑音性評価

耐雑音性の評価は，TUSK [31] により実施することとした．TUSK は，F0 が既知の信号からテスト用の波形を生成するため，EGG から F0 の真値を比較する音声データベースを用いた評価と比較して，F0 の真値を確実に与えることができる利点がある．TUSK では 6 つの尺度を用いて評価することが可能であるが，ここでは耐雑音性の評価に対応する ACT 5 のみ実施する．ACT 5 では，F0 軌跡 $f_0(t)$ と生成される評価用信号 $x(t)$ との関係が以下の式で表される．

$$x(t) = n(t) + \sum_{k=1}^{K} \cos\left(2\pi k \int_0^t f_0(\tau) d\tau\right), \tag{4}$$

ここで，$n(t)$ は加算される雑音，K は調波の数を示す．N は，ナイキスト周波数を超えない範囲での最大値に設定される．

真値となる F0 軌跡 $f_0(t)$ は，Klatt により提案された揺らぎ [38] を含む以下の式で表される．

$$f_0(t) = f + \frac{f}{200}(\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t)), \tag{5}$$

f は，基準となる F0 であり，生成される F0 軌跡の平均に対応する．

ここでは，サンプリング周波数は 48 kHz，雑音 $n(t)$ をホワイトノイズとして，SNR を 0 から 60 dB まで 0.5 dB 刻みで変化させて評価した．f は TUSK を提案する論文と同様に，440 Hz とした．雑音に対する依存性を勘案し，同 SNR で 100 回試行してその中央値を最終的な結果として採用した．

図 4 に評価結果を示す．図の横軸は SNR，縦軸は真値との RMS error に対応する．提案する Harvest は，YIN 以外の方法と比較すると，同 SNR における推定性能は優れていることが確認できる．一方，SNR が 7 dB を下回ると YIN の推定性能のほうが良好であることも確認できる．

5. 考　　察

本実験結果は，Harvest が期待通り高い性能と耐雑音性を有することを示した．ここでは，特に音声データベースを用いた評価における EGG 信号の問題点と，両評価を含めた Harvest の有効性について議論する．

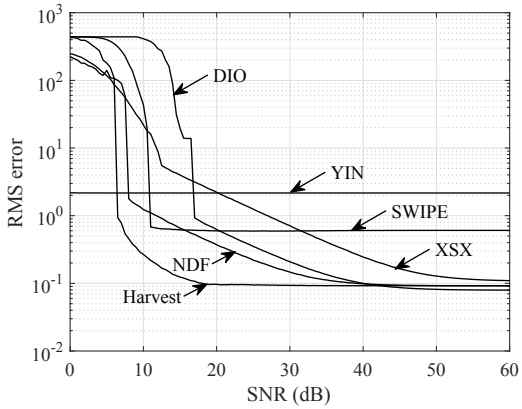

図 4　人工的に生成された信号を用いた耐雑音性評価の結果.

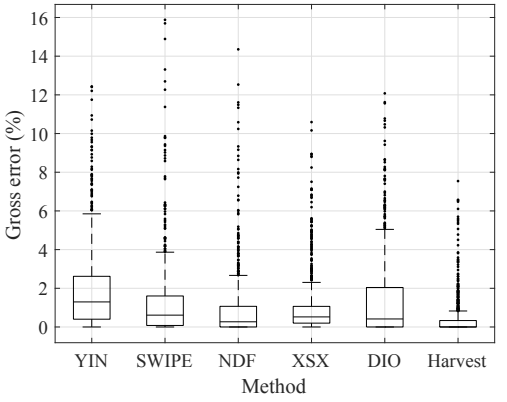

図 5　840 の音声から個別に求めた Gross error に基づく箱ひげ図.

5.1　音声データベースを利用した評価結果と問題点

図 5 は，実験に用いた 840 の音声について得られた Gross error から計算した箱ひげ図を示す．どの手法に関しても，一部の音声の Gross error が極端に高いことが確認できる．この理由として，EGG 信号からの F0 推定は NDF を用い，VUV 判定はデータベースに収録された結果を用いたことが挙げられる．例えば，図 6 は，全ての手法で Gross error が高い音声の波形と NDF により求められた F0 の真値を示す．VUV 区間は，音声データベースに収録されている時系列をそのまま採用した．1870 ms から 2050 ms は波形から目視で判断すると無声区間であるが，音声データベースでは有声音と判定していることが確認できる．このような VUV 判定のミスは他の音声でもしばしば生じているため，本実験で Gross error を計算する分母となるフレーム数には無声音区間が入っているといえる．

現在，F0 推定法に関する論文の多くは音声データベースを用いて僅かな差で優劣を議論しているが，真値の信頼性の観点から，微差の議論は意味をなさない可能性がある．信頼できる区間のみ抽出して評価を行うことでこれらの問題は回避可能であるが，例えば音声合成が目的である場合は，最終的に高品質な音声を合成できる F0 軌跡であることが重要である．Harvest は最良の結果を示したが，今後は，微差ではなく分析合成音の評価も合わせて行うことが，音声合成を目指した F0 推定法の優劣の議論には不可欠であるといえる．

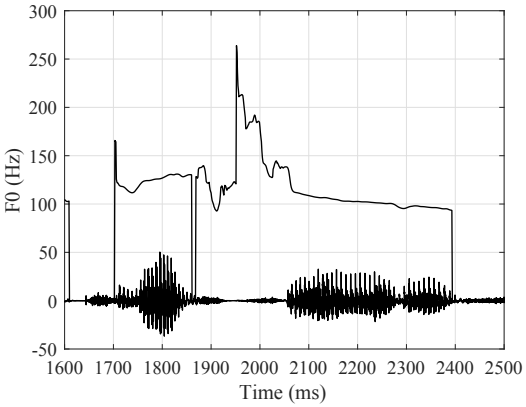

図 6　音声データベースの VUV 判定が誤っている例．1885 ms から 2050 ms は目視で判定すると無声区間であるが，データベースでは有声音として判定されている．NDF は無音区間に F0 を強制的に当てはめることが可能である．

5.2　耐雑音性に関する考察

TUSK ACT 5 による評価の結果，耐雑音性については，Harvest は NDF と比較しても良好な性能を示すことが確認された．TUSK は真値が既知な信号を生成して評価に利用するため，音声データベースの評価での問題は原理的に生じない利点がある．本評価では，音声データベースと TUSK による評価を並列して実施することで，F0 推定法が性能と耐雑音性の両面で優れていることを示した．

F0 推定法の評価は，どのような目的の F0 推定法であるかにより評価法を吟味することが重要である．例えば，音声分析合成の場合，無声音を有声音と誤推定した場合でも，非周期性指標推定が適切に動作すれば影響を吸収できるが，逆の場合は大きな品質の劣化となる．本実験は，耐雑音性について Harvest は SNR が 7 dB 以下で YIN よりも劣ることを示したが，高品質音声合成の観点から必要となる耐雑音性能については，今後検討する必要がある．

6.　おわりに

本稿では，F0 推定法 Harvest を提案して有効性を示すとともに，EGG 信号を用いた実音声の評価の問題点についても議論した．Harvest は，多チャネルに信号を分解し，1 フレームの推定について多数の FFT を行う高い計算コストを要求する一方，高い耐雑音性と推定性能を両立することに成功した．

現在の Harvest は，スペクトル上の局所的なピークを全て F0 候補とする仕様であるため，不要なピークを除去することがさらなる精度改善に繋がる．合成音声の品質の観点から F0 推定法を評価することも重要な課題といえる．多チャネルの信号処理は GPU を用いた並列処理を相性が良いため，高速化を目的とした本手法の GPU 版の実装にも取り組む計画である．

謝　　辞

本研究は，科研費 JP15H02726, JP16H05899, JP16H01734 の支援を受けて実施された．

文　献

[1] H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol.11, no.2, pp.169–177, 1939.

[2] M.V. Mathews, J.E. Miller, and E.E. David, “Pitch synchronous analysis of voiced sounds,” J. Acoust. Soc. Am., vol.33, pp.179–186, 1961.

[3] M. Morise, “CheapTrick, a spectral envelope estimator for high-quality speech synthesis,” Speech Communication, vol.67, pp.1–7, 2015.

[4] M. Morise, “Error evaluation of an f0-adaptive spectral envelope estimator in robustness against the additive noise and f0 error,” IEICE Trans. Inf. & Syst., vol.E98-D, no.7, pp.1405–1408, 2015.

[5] T. Nakano and M. Goto, “A spectral envelope estimation method based on f0-adaptive multi-frame integration analysis,” in Proc. SAPA-SCALE 2012, pp.11–16, 2012.

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

[7] A. Lee, S. Furuya, M. Morise, P. Iltis, and E. Altenmüller, “Quantification of instability of tone production in embouchure dystonia,” Parkinsonism & related disorders, vol.20, pp.179–186, 2014.

[8] M. Morise, H. Kawahara, and H. Katayose, “Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech,” in Proc. AES 35th International Conference, CD-ROM, pp.CD–ROM, 2009.

[9] 森勢将雅，河原英紀，西浦敬信，“基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法，” 電子情報通信学会 論文誌 D，vol.J93-D，no.2，pp.109–117，2010.

[10] H. Zen, K. Tokuda, and A.W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol.51, pp.1039–1064, 2009.

[11] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. & Syst., vol.E99-D, pp.1877–1884, 2016.

[12] Z. Wu, O. Watts, and S. King, “Merlin: An open source neural network speech synthesis system,” in Proc. of SSW 2016, pp.218–223, 2016.

[13] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, “Average magnitude difference function pitch extractor,” IEEE Transactions on acoustic, speech, and signal processing, vol.ASSP-22, no.5, pp.353–362, 1974.

[14] A.M. Noll, “Short-time spectrum and “cepstrum” techniques for vocal pitch detection,” J. Acoust. Soc. Am., vol.36, no.2, pp.269–302, 1964.

[15] A.M. Noll, “Cepstrum pitch determination,” J. Acoust. Soc. Am., vol.41, no.2, pp.293–309, 1967.

[16] A. Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” J. Acoust. Soc. Am., vol.111, no.4, pp.1917–1930, 2002.

[17] M. Mauch and S. Dixon, “PYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in Proc. ICASSP2014, pp.659–663, 2014.

[18] A. Camacho and J.G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” J. Acoust. Soc. Am., vol.124, no.3, pp.1638–1652, 2008.

[19] 佐宗晃，中村尚五，“ウェーブレット変換を用いたピッチ抽出の一方法，” 電子情報通信学会論文誌 A，vol.J80-A，no.11，pp.1848–1856，1997.

[20] B. Yegnanarayana and K.S.R. Murty, “Event-based instantaneous fundamental frequency estimation from speech signals,” IEEE Transactions on Audio, Speech, and Language Processing, vol.17, no.4, pp.614–624, 2009.

[21] H. Kawahara, T. Irino, and M. Morise, “An interference-free representation of instantaneous frequency of periodic signals and its application to f0 extraction,” in Proc. of ICASSP 2011, pp.5420–5423, 2011.

[22] T. Shimamura and H. Kobayashi, “Weighted autocorrelation for pitch extraction of noisy speech,” IEEE Transactions on speech and audio processing, vol.9, no.7, pp.727–730, 2001.

[23] T. Nakatani and T. Irino, “Robust and accurate fundamental frequency estimation based on dominant harmonic components,” J. Acoust. Soc. Am., vol.116, no.6, pp.3690–3700, 2004.

[24] 大村浩，田中和世，“基本波フィルタリング法による精細ピッチパターンの抽出，” 日本音響学会誌，vol.51，no.7，pp.509–518，1995.

[25] R. Daido and Y. Hisaminato, “A fast and accurate fundamental frequency estimator using recursive moving average filters,” in Proc. INTERSPEECH 2016, pp.2160–2164, 2016.

[26] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, “Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution,” in Proc. ICASSP 2013, pp.6796–6801, 2013.

[27] H. Kawahara, Y. Agiomyrgiannakis, and H. Zen, “Using instantaneous frequency and aperiodicity detection to estimate f0 for high-quality speech synthesis,” arXiv preprint arXiv:1605.07809, 2016.

[28] A.H. Nuttall, “Some windows with very good sidelobe behavior,” IEEE Trans. on acoust., speech, and signal processing, vol.29, no.1, pp.84–91, 1981.

[29] J.L. Flanagan and R.M. Golden, “Phase vocoder,” The Bell System Technical Journal, vol.45, no.9, pp.1493–1509, 2009.

[30] 森勢将雅，“基本波検出に基づく f0 推定法の耐雑音性向上，” 情報処理学会音声言語情報処理研究会，vol.2016-SLP-110，no.5，pp.1–6，2016.

[31] M. Morise and H. Kawhahara, “TUSK: A framework for overviewing the performance of f0 estimators,” in Proc. INTERSPEECH 2016, pp.1790–1794, 2016.

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

[33] H. Kawahara, A. Cheveigné, H. Banno, T. Takahashi, and T. Irino, “Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight,” in Proc. Interspeech2005, pp.537–540, 2005.

[34] 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,” in Proc. ICASSP2008, pp.3933–3936, 2008.

[35] 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, 2011.

[36] 阿竹義徳，入野俊夫，河原英紀，陸金林，中村哲，鹿野清宏，“調波成分の瞬時周波数を用いた基本周波数推定方法，” 電子情報通信学会論文誌 D，vol.J83-DII，no.11，pp.2077–2086，2000.

[37] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal, “A comparative performance study of several pitch detection algorithms,” IEEE Transactions on acoustic, speech, and signal processing, vol.ASSP-24, no.5, pp.399–418, 1976.

[38] D.H. Klatt and L.C. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am., vol.82, no.2, pp.820–857, 1990.