Vol.2016-MUS-110 No.1
2016/2/29

基本周波数推定法の性能を概観するフレームワークの試作

森勢 将雅[1,a)]　河原 英紀[2]

概要：本稿では，基本周波数（F0）推定法の性能を概観するためのフレームワーク TUSK を提案し，計算機シミュレーションにより性能の確認を行う．F0 推定法は，時間波形の特徴やスペクトルの特徴に基づく方法など，多数の方法がすでに提案されている．それぞれの方法は，音声データベースに収録された音声により性能を評価し有効性が検証されている一方，データベースの音声にも F0 や音色の偏りが存在することから，正確な優劣を議論することは容易ではない．本研究では，F0 推定法にはそれぞれ性能を発揮する特性の音声や収録環境があるという仮説のもと，優劣をつけるのではなく性能を概観し，ユーザが F0 推定法を選択する手がかりを与えることを目指す．提案する TUSK では，音声・歌声分析に必要な，ビブラートや耐雑音性などの項目について人工的に生成した音源による評価を実施し，各項目についての性能を計測することを可能にする．本稿では，TUSK のコンセプトと各評価項目について示し，近年提案された高性能な F0 推定法を用いた比較評価により有効性を示す．

キーワード：音声分析，基本周波数，時間変動，耐雑音性

Prototype of a framework for overviewing the performance of F0 estimators

Morise Masanori[1,a)]　Kawahara Hideki[2]

Abstract: This article represents a framework for overviewing the performance of fundamental frequency (F0) estimator and evaluates its effectiveness. Many F0 estimators have been proposed, and their effectivenesses have been evaluated by speech databases. On the other hand, since the evaluation result depends on the speech database used for the evaluation, it is difficult to fairly evaluate the estimators. The framework, named TUSK, does not rank the estimators but attempts to overview them. In this article, we introduce the concept of TUSK and evaluation criteria, and its effectiveness is examined by several modern F0 estimators.

Keywords: Speech analysis, fundamental frequency, temporal variation, noise robustness

1. はじめに

音声や歌声合成のソフトウェアが広く一般で利用されるようになり，それに伴い音声分析・合成を行う技術への需要も高まっている．音声の分析において，基本周波数 (F0) は音声を構成する主要なパラメータの 1 つであり，高い精度での F0 推定法の実現は，現在に至るまで継続して進められている研究テーマである．あらゆる音声から F0 を高精度に推定する方法が提案されれば，音声分析，合成を含む幅広い研究分野に対して大きな貢献となる．しかしながら，音声の F0 は時間とともに変動するため急激に変化する F0 の推定は困難であり，また，収録環境によっては雑音や残響の混入という多くの問題に対処する必要がある．特定の問題に対処することは，別の問題での誤差を拡大する可能性もあり，全ての音声から F0 を高精度に推定する方法は未だに確立していない．

Vocoder [1] 等分析合成系に関する音声合成の分野では，

[1] 山梨大学
University of Yamanashi, 4-3-11, Takeda, Kofu, 400–8511, Japan
[2] 和歌山大学
Wakayama University, 930, Sakaedani, Wakayama, 640–8510, Japan.
[a)] mmorise@yamanashi.ac.jp

音声はある程度高 SNR の環境で収録されることが期待される．音声の音色の相当するスペクトル包絡推定では，音声の F0 情報に基づいて切り出す窓関数を調整することにより高い性能を達成できることが示されていることから [2], [3], [4]，F0 の時系列を精密に推定する方法が必要となる．一方，実環境で収録した音声を用いる分野では，雑音や残響が混入した音声が分析対象となるため，厳密な推定精度よりもそれら要因に頑健であることが期待される．F0 推定法は目的に応じて様々な方法が提案されているが，万能な方法が無い以上，使用者は，分析する音声に対して適切な分析法を選択しなければならない．この選択を助けるには，音声データベースを用いた推定精度で優劣を競うことより，音声の特徴から適切な F0 推定法を見出す手がかりを与えることが重要である．

本研究では，F0 推定法の優劣を示すためではなく，F0 推定法の特性を概観することを狙った評価フレームワーク「TUSK」を提案することで，この問題の解決を図る．任意の音声データベースを利用する場合，F0 の分布や話者性など音色に偏りが生じるため，TUSK では，人工的に生成した信号を利用する．F0 推定法を選択する手がかりとして 6 つの評価項目を提案し，各評価項目に対応するパラメータを段階的に変化させつつ推定性能を計測することで，各項目における評価を可能にする．本稿では，提案する TUSK のコンセプトと 6 つの評価項目を述べ，最新の方法を含むいくつかの F0 推定法を用いた実験によりその有効性を検証する．

2. F0 推定法の種類と性能の評価法

F0 推定法は古くから研究されており [5]，時間波形における周期性に着目した方法やパワースペクトルの特徴に着目した方法など数多くの方法が提案されている．波形の周期性については，相関を用いた方法 [6] が古くから検討されており，性能を向上させるための理論を加えた YIN[7]，さらに改良を加えた pYIN[8] が提案されている．パワースペクトルの特徴に関しては，古くは Cepstrum [9], [10] が提案されており，近年では，スペクトル構造に着目した高精度な方法として SWIPE′ が提案されている [11]．その他にも，瞬時周波数を用いた方法 [12]，ウェーブレット変換を用いた方法 [13]，声帯振動のイベントを検出しそこから F0 を推定する方法 [14]，周期信号は調波複合音で表現できるため F0 に対応する基本波の周波数を直接求める方法 [15] や，それを改良した方法として DIO[16] が存在する．新たな特徴量を提案する研究だけではなく，耐雑音性に特化した改良についても検討されている [17], [18].

これら手法の評価は，一般的に，人間の発話した音声を対象に行われてきた．評価用の音声データベースとして，日本語であれば阿竹らによる 14 名の発話音声から構成されるデータベース [12]，英語であれば CMU のデータベース [*1] や Bagshaw のデータベース [*2] などが用いられる．これらの音声データベースには，声帯の開閉の情報に相当する Electrogastrogram (EGG) 信号が同時に収録されており，EGG から求めた F0 を真値として音声波形の F0 推定の精度を評価する．評価指標についても，古くは Fine pitch error (FPA)，Gross pitch error (GPA) [19] や Gross error [7] が提案され，現在でも用いられている．

F0 推定法の評価をこれらの音声データベースと評価指標で行う場合，特定の音声データベースでは高性能だが別の音声データベースでは他の方法が優れているケースや，同じデータベースでも Gross Error では優れていても FPE では劣るケースが生じる．音声データベースに関しては，データベースごとに収録話者の F0 の分布や収録環境が異なることが原因となり，特定の音声においての優劣のみ評価しており汎用性に問題が生じる可能性がある．歌声の分析では F0 の範囲が広く，ビブラートやポルタメントなど F0 を短時間で変動させる技巧も存在するため，分析音声の F0 軌跡の急峻な変化をどの程度追従可能であるかなども検証する必要がある．

3. TUSK のコンセプトと評価項目

TUSK では，音声・歌声の F0 推定で重要になると想定する 6 つの項目を提案し，それらの項目に対応するパラメータを段階的に変化させた模擬音声を対象に各 F0 推定法の評価を行うことで，各項目についての性能の把握を目指す．評価用信号は，人間の音声ではなく F0 軌跡から生成した調波複合音とすることで，真値を既知として誤差評価することが可能となる．

3.1 共通する評価用信号と誤差評価指標

まず，全評価で共通して用いる調波複合音を，以下の式により定義する．

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

ここで $n(t)$ は加算性の雑音，$h(t)$ はインパルス応答，$*$ は畳み込みを表す記号，$f_0(t)$ は F0 の時系列を示す．a_k と θ_k は，k 番目の調波の振幅と位相をそれぞれ示す．K は調波数に対応し，$Kf_0(t)$ が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．TUSK では，調波複合音における各パラメータを段階的に変化させることで，特定の項目が性能に与える影響を評価することが可能となる．

実験に用いる F0 軌跡 $f_0(t)$ は，基本的にターゲットとなる基本周波数 f_c で固定される．ただし，人間の音声には揺らぎが含まれるため，F0 軌跡は完全な固定値にするのではなく，Klatt により提案された，以下の式により定義

[*1] http://festvox.org/cmu_arctic/index.html
[*2] http://www.cstr.ed.ac.uk/research/projects/fda/

される揺らぎ成分 [20] を加える．

$$\Delta f_0(t) = \frac{FL}{50}\frac{f_c}{100}(\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t)), \quad (1)$$

FL はフラッターに相当するパラメータであり，文献 [20] に倣い 25 で固定する．ベースとなる F0 軌跡は，$f_0(t) = f_c$ とし，このデルタ成分を加えることで最終的な F0 軌跡とする．なお，基本パラメータは，$n(t) = 0$，$h(t) = \delta(t)$，$a_k = 0$，$\theta_k = 1$，$f_c = 440$ とし，各評価ではこれらのパラメータのうち 1 つを段階的に操作する．

F0 の推定精度を示す指標は複数存在するが，TUSK では，真値と推定値からシンプルな RMS 誤差を求めることとする．一般的に，真値の倍や半分の値を誤推定する問題が知られており，従来の指標には，そのような大幅な誤差を計測する意図がある．例えば Gross error では，真値の ±20% に推定値が含まれるか否かを評価する指標だが，20%以内であれば全て正解とみなされるため，F0 軌跡をどの程度精密に推定可能であるか評価することはできない．FPE や GPE は，誤差が閾値以内に収まる比率と，閾値以内に収まった F0 が真値からどの程度離れているかを評価する．Gross error より精密な評価が行えるが，これも閾値の設定により推定結果が変化することが考えられるため，TUSK ではより簡略化された評価のため F0 軌跡の RMS 誤差を用いる．

3.2 ACT1: 高さと推定精度との関係

まず，F0 の高低と推定精度との関係を確認するため，F0 の高低をパラメータとした評価を行う．この調波複合音の F0 には Klatt により提案された揺らぎしか存在せず，信号には雑音や残響を一切含まないため，多くの F0 推定法ではこの評価での結果が最良の精度となる．ベースとなる基本周波数 f_c の周波数レンジを任意に与えることで，低域から高域までの F0 をどの程度正確に推定可能であるか検証可能となる．分析対象となる音声が男性か女性かなど高さが明らかである場合，この評価結果によりある程度結果を予測することが可能となる．

3.3 ACT2: F0 軌跡の時間変化に対する頑健性

多くの F0 推定法は，声帯振動が周期的に生じている前提で理論が組まれているが，実音声の F0 は時間とともに変化する．特に，歌声分析では，ビブラートやポルタメントのように，F0 が急峻に変化する場合でも適切な F0 を求めることが望まれる．F0 推定では短い区間において定常性を仮定することとなるが，F0 軌跡の時間変化が急峻である場合，適切に推定できない可能性が考えられる．ACT2 では，F0 軌跡にビブラート的な振動成分を与えることで，F0 の時間変化に対する頑健性を評価する．時間変動成分は，以下の式により与えられる．

$$f_v(t) = \sqrt{\alpha f_c}\cos\left(\sqrt{\alpha f_c}t\right), \quad (2)$$

α は F0 の時間変動の強さを表し，最大の傾斜が αf_c になる特徴を有する．評価に用いる F0 軌跡は，ベースとなる f_c に，$\Delta f_0(t)$ とビブラート強度 $f_v(t)$ が加算されたものとなる．ACT2 では，ビブラート強度を示すパラメータ α と推定精度との関係性により，F0 の時間変化が推定結果に与える影響を評価する．

3.4 ACT3: 調波複合音の振幅のダイナミックレンジが性能に与える影響

各調波の振幅 a_k が固定された信号は実音声とは大きく特性が異なる．ただし，実音声も模擬する振幅のデザインは容易ではないため，ACT3 では，式 (1) における a_k をランダマイズすることで，調波構造の乱れが推定結果に与える影響を計測することを目指す．一部の F0 推定法は，パワースペクトルの調波構造や基本波の周波数に着目して F0 を推定するため，振幅のランダマイズの幅を制御パラメータとすることにより，系統的な変化として評価することが可能となる．TUSK では，対数軸上で任意のダイナミックレンジを与えるように，一様乱数で a_k を決定する．隣接する調波の振幅差はランダムにより揺らぐこととなるため，調波構造に基づく方法は影響されることや，基本波の周波数を推定する方法も推定精度が影響されることが予想される．

3.5 ACT4: 調波複合音の位相のランダマイズが性能に与える影響

ACT3 では，各調波の振幅差が推定結果に与える影響を計測し，ACT4 では，各調波の位相差が推定結果に与える影響を計測する．式 (1) における θ_k をランダマイズすることで，調波構造の乱れが推定結果に与える影響を評価する．なお，ACT4 においては，ランダマイズのダイナミックレンジがパラメータであり，最大値は 2π でランダマイズには一様乱数を用いる．

3.6 ACT5: 耐雑音性

実音声は収録環境に依存した雑音が混入するため，耐雑音性に関する評価は，収録音声が高 SNR であることが保証される場合を除き必要不可欠である．ACT5 では，$n(t)$ に任意の雑音を任意の SNR で与えることにより，SNR と推定精度との関係を計測することが目指す．雑音の種類については定めていないが，全帯域での SNR を固定する場合はホワイトノイズが望ましく，実環境を想定する場合はピンクノイズを用いることが望ましいといえる．この 2 つの雑音から SNR と推定精度との関係を観察することにより，耐雑音性についての特性を評価する．

3.7 ACT6: 残響に対する頑健性

最後に，残響が計測結果に与える影響を調査するための評価を ACT6 として行う．まず，残響に相当するインパルス応答 $h(t)$ の時間エンベロープを，以下の式により定義する．

$$h_e(t) = \begin{cases} 0 & (t < 0) \\ 1 & (t = 0) \\ \dfrac{\exp\left(\dfrac{t\log(0.001)}{r}\right)}{\sqrt{10}} & (otherwise) \end{cases} \quad (3)$$

ここで，r は残響時間に相当するパラメータである．t が 0 より大きい場合について $h_e(t)$ と無相関雑音との積を計算することで，最終的なインパルス応答 $h(t)$ とする．TUSK では，1 つのパラメータで 1 項目を評価する狙いがあるため，本インパルス応答には初期反射音を含まず，直接音の直後に残響成分が到来するよう設計されている．減衰開始時刻の振幅は，EDT (Early Decay Time) が 10 dB 減衰するまでの時刻から求める指標であることに着目し，初期反射音を含まない（EDT が 0 になる）条件での最大振幅として設定されている．この設計により，残響を操作するパラメータは r のみとすることが可能となる．

4. 評価

提案する TUSK を用いて，いくつかの F0 推定法の性能を評価する．F0 推定法には，波形の相関に基づく推定法として YIN[7], スペクトルに着目した方法として SWIPE′[11]（以下では単に SWIPE とする）を用いる．他にも高品質な音声合成に向けた方法として，筆者が提案する DIO[16], DIO により推定された結果を瞬時周波数により補正した DIO+StoneMask, STRAIGHT で利用される NDF[21], TANDEM-STRAIGHT で利用される XSX[22], [23] を利用する．YIN と SWIPE については，論文の著者が Matlab のソースコードを公開しているため，それらを利用した．DIO と StoneMask については，以下の Web サイトから入手可能である [*3]．なお，各プログラムには F0 を探索する範囲が設定されているため，事前に実験で用いる周波数範囲をカバーするよう下限を 40，上限 1000 Hz に設定している．

評価に用いる信号長は 1.2 s とし，$f_0(t)$ は 1 ms 毎に求める．評価には 0.1 から 1.1 s の区間で得られる 1000 サンプルの結果を利用し，真値からの差分に基づいて RMS 誤差を計算する．これは，波形の開始・終了時における F0 の定義は困難であり，窓関数により切り出す際の処理など，実装の工夫が推定結果に与える影響を防ぐ狙いがある．評価用信号は，サンプリング周波数を 48 kHz とする．ACT1 から ACT6 までの各評価に用いるパラメータは表 1 に示す通りである．ACT5 については，ホワイトノイズとピンクノイズの 2 種類の雑音により評価する．なお，ランダム性の影響を除外するため，ACT3 から ACT6 は，同条件で 100 回試行し，結果の中央値を採用することとした．

[*3] http://ml.cs.yamanashi.ac.jp/world/

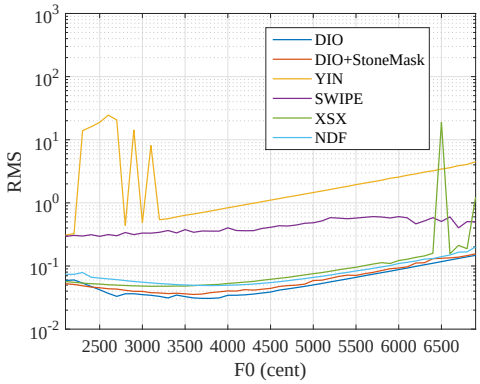

図 1 TUSK ACT1 の評価結果

表 1 各項目のパラメータ

実験項目	パラメータ
ACT1	f_c: 2100...6900 cent (A1...A5)
ACT2	α: 0...25
ACT3	a_k: 0...40 dB
ACT4	θ_k: 0...2π
ACT5 (white noise)	SNR: 0...60 dB
ACT5 (pink noise)	SNR: 0...60 dB
ACT6	r: 10...1000 ms

4.1 実験結果：ACT1

始めに，F0 と推定誤差との関係を図 1 に基づいて議論する．図の横軸は cent 表記の F0 であり，下限が 2100 cent (55 Hz)，上限が 6900 cent (880 Hz) である．推定性能は YIN が明らかに低く，F0 が高くなるにつれて誤差が増加傾向にある．また，低い F0 に関しては真値とは全く異なる推定値を返していることが分かる．SWIPE は，F0 と誤差に比例関係はないが，他の手法と比較すると相対的に性能が低いといえる．YIN や XSX が特定の高さにおいて誤差が大幅に増加しているのは，半分や倍の F0 として推定されたか，あるいは無声区間と判定されたことを示す．雑音が含まれない音声の場合，残りの方法は概ね同程度の誤差となることも観測できる．

4.2 実験結果：ACT2

図 2 は，F0 の時間変動としてビブラート強度 α と推定誤差との関係を示す．全ての方法でビブラート強度が増えることで誤差が増えることが確認できる．ただし，SWIPE のみは，α が 20 を超えたあたりで誤差が急激に増えていることが確認できる．つまり，音声の F0 変動については，極端に早い変動が含まれる場合は SWIPE は適さない，つまり SWIPE は時間分解能に劣ることが示唆される．

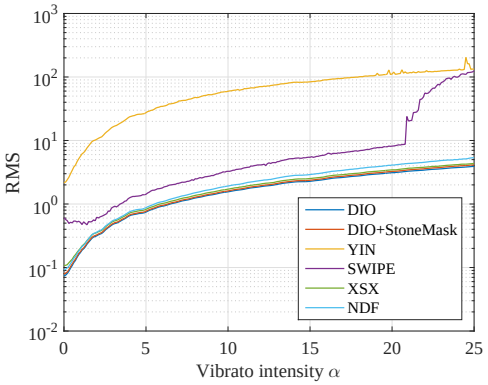

図 2 TUSK ACT2 の評価結果

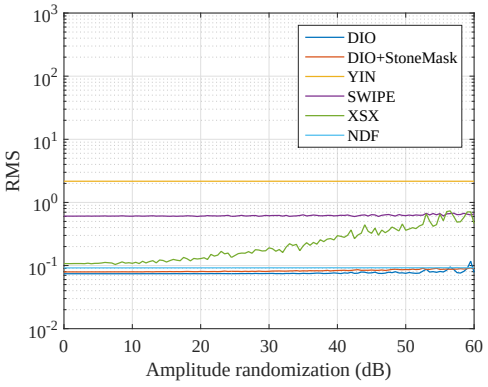

図 3 TUSK ACT3 の評価結果

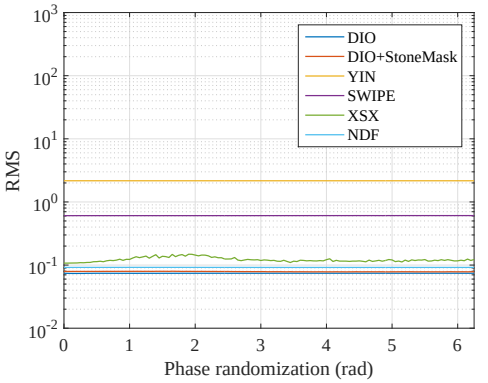

図 4 TUSK ACT4 の評価結果

4.3 実験結果：ACT3, ACT4

ACT3, ACT4 についての結果を図 3,4 に示す．振幅に関しては，XSX がダイナミックレンジの向上とともに誤差が増加する傾向がある．

位相に関しては，XSX 以外の方法は影響されないといえる．これは，DIO はローパスフィルタにより基本波のみを取り出す方法のため，また SWIPE などもパワースペクトルの特性に基づくため影響がほぼ生じないといえる．ただし XSX についても 0.1 Hz 以下の範囲での揺れである．

4.4 実験結果：ACT5

ACT5 についての結果を図 5,6 に示す．どちらも SNR が低下するについて誤差が拡大するが，YIN や SWIPE は

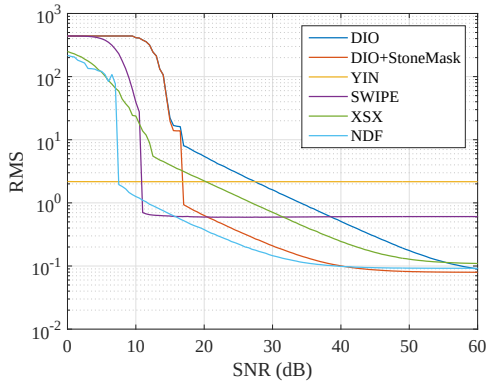

図 5 TUSK ACT5 (whilte noise) の評価結果

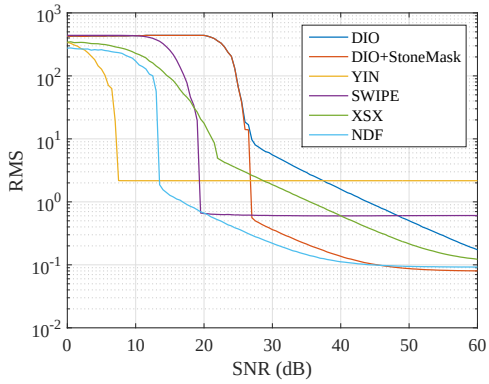

図 6 TUSK ACT5 (pink noise) の評価結果

雑音に影響されにくい傾向が確認できる．ホワイトノイズとピンクノイズとの差については，ピンクノイズのほうが低域にパワーが集中するため，基本波を取り出す DIO や低域の調波構造を利用する XSX では，より高い SNR でも誤差が拡大する傾向にある．DIO+StoneMask は，ACT1 では DIO よりも誤差を増大させていたが，耐雑音性の観点からは有効な手段であることが確認できる．

4.5 実験結果：ACT6

耐残響性に関する評価結果を図 7 に示す．横軸は残響時間だが，残響は，インパルス応答の波形エンベロープが直接音より 10 dB 小さい振幅から始まり，初期反射音も存在しないため，一般的な残響時間とは対応しないことに注意が必要である．YIN のみ，特定の残響時間まで誤差が減少する特性が見えるが，YIN 以外の方法は，残響時間の増加に伴い誤差が増大している傾向が確認できる．

5. 考察

本実験により，各方法がどのような特性を有するか把握することができたといえる．例えば，YIN はクリーンな音声の場合他の方法よりも誤差が大きいが，耐雑音性と耐残響性には優れていること，DIO は高 SNR の音声に対して高い性能を発揮するといえる．一方，基本波を取り出す方法であるため，雑音に対する影響を強く受けるともいえる．

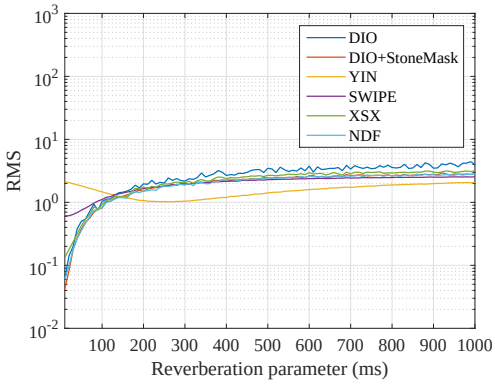

図 7 TUSK ACT6 の評価結果

STRAIGHT で利用される NDF に関しては，全ての条件でバランス良く F0 を推定可能であることが確認できる．SWIPE は，YIN と NDF の中間的な特性を有する．これらの評価結果に基づき，使用者は，音声の収録環境などから，適切な方法を選択可能であると考えられる．

6. おわりに

本稿では，F0 推定法の性能を概観するためのフレームワーク TUSK を提案し，近年提案された高精度な F0 推定法の性能評価により有効性を検証した．その結果，各手法の性能を適切に比較可能であることが示された．使用者は，分析対象となる音声の特性から適切な方法を選択することが可能になることが期待される．

今後は，これらの評価結果が実音声を分析する際に適切に対応しているかどうか，多数の音声データベースを用いて検証する必要がある．また，音声から環境を推定し，適切な F0 推定法を推薦するシステムを開発することで，音声分析時の手間を改善する方法について取り組みたい．

謝辞 本研究は，科研費 15H02726，26540087，および東北大学電気通信研究所 共同プロジェクト（H25/A08）の支援を受けて実施された．

参考文献

[1] Dudley, H.: Remaking speech, *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177 (1939).
[2] Morise, M.: CheapTrick, a spectral envelope estimator for high-quality speech synthesis, *Speech Communication*, Vol. 67, pp. 1–7 (2015).
[3] Morise, M.: 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).
[4] Nakano, T. and Goto, M.: A spectral envelope estimation method based on F0-adaptive multi-frame integration analysis, *in Proc. SAPA-SCALE 2012*, pp. 11–16 (2012).
[5] Hess, W.: *Pitch determination of speech signals*, Springer-Verlag (1983).
[6] Ross, M., Shaffer, H., Cohen, A., Freudberg, R. and Manley, H.: Average magnitude difference function pitch extractor, *IEEE Transactions on acoustic, speech, and signal processing*, Vol. ASSP-22, No. 5, pp. 353–362 (1974).
[7] Cheveigné, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music, *J. Acoust. Soc. Am.*, Vol. 111, No. 4, pp. 1917–1930 (2002).
[8] Mauch, M. and Dixon, S.: PYIN: A fundamental frequency estimator using probabilistic threshold distributions, *in Proc. ICASSP2014*, pp. 659–663 (2014).
[9] Noll, A.: Short-time spectrum and “cepstrum” techniques for vocal pitch detection, *J. Acoust. Soc. Am.*, Vol. 36, No. 2, pp. 269–302 (1964).
[10] Noll, A.: Cepstrum pitch determination, *J. Acoust. Soc. Am.*, Vol. 41, No. 2, pp. 293–309 (1967).
[11] Camacho, A. and Harris, J. G.: A sawtooth waveform inspired pitch estimator for speech and music, *J. Acoust. Soc. Am.*, Vol. 124, No. 3, pp. 1638–1652 (2008).
[12] 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).
[13] 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).
[14] Yegnanarayana, B. and Murty, K.: 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).
[15] 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).
[16] 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).
[17] Shimamura, T. and Kobayashi, H.: Weighted autocorrelation for pitch extraction of noisy speech, *IEEE Transactions on speech and audio processing*, Vol. 9, No. 7, pp. 727–730 (2001).
[18] Nakatani, T. and Irino, T.: Robust and accurate fundamental frequency estimation based on dominant harmonic components, *J. Acoust. Soc. Am.*, Vol. 116, No. 6, pp. 3690–3700 (2004).
[19] Rabiner, L., Cheng, M., Rosenberg, A. and McGonegal, C.: 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).
[20] Klatt, D. and Klatt, L.: 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).
[21] Kawahara, H., Cheveigné, A., Banno, H., Takahashi, T. and Irino, T.: Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT, *in Proc. Interspeech2005*, pp. 537–540 (2005).
[22] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: 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).
[23] Kawahara, H. and Morise, M.: 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).