Vol.2014-DCC-8 No.8
Vol.2014-MUS-105 No.8
2014/11/21

高品質な歌声・音声合成を目的としたスペクトル包絡推定法 CheapTrickの誤差評価

森勢 将雅[1,a)]

概要：筆者は Vocoder 型の歌声・音声合成技術に関する研究に取り組んでおり，本稿では新たに提案したスペクトル包絡推定法の誤差評価について述べる．CheapTrick と命名した提案法は，精密に調整された 1 つの窓関数による波形の切り出しとスペクトルの平滑化により，高精度なスペクトル包絡推定を実現する．過去の検討では，提案法と従来法とで合成音声の品質に関する主観評価を実施し，CheapTrick は最高品質の音声が合成可能な利点を明らかにした．CheapTrick は基本周波数 (F0) に基づいて窓関数を設計する必要があるため，音声から推定された F0 に誤差が含まれる場合推定結果にも影響する．また，実音声には雑音成分が混入するため，雑音耐性についても評価が必要となる．本稿では，耐雑音性，および F0 の推定誤差が推定結果に与える影響について，客観指標を用いた誤差評価の結果に基づいて検討する．評価の結果，CheapTrick は，どちらの耐性についても実用的な条件下において従来法より高く，優れた性能を有することが示された．

1. はじめに

高品質音声合成は，いまや音声処理における主要な研究領域の 1 つとなり，音声だけではなく表情豊かな歌声合成技術はコンシューマ向けの製品として実装され販売されている．音声合成方式にはいくつもの種類があり，例えば PSOLA [1]，sinusoidal model [2] や phase vocoder などが知られている．本稿では，Vocoder [3] の考え，すなわち，音声を基本周波数 (F0) と調音フィルタとに分解し合成する方式を取り扱う．我々が提案した最新の音声分析合成方式は，音声を F0，スペクトル包絡，非周期性指標に分解するアルゴリズム，および 3 つのパラメータから音声を再合成するアルゴリズムから構成される．とりわけ，スペクトル包絡は，音韻や個人性など高さ以外の多様な情報の認識に役立つことから，音声認識，合成，知覚など多くの研究領域で利用されており，それぞれの研究領域で求められる条件を満足する様々な方法が提案されてきた [4], [5], [6], [7], [8], [9]．高品質音声合成を目指したスペクトル包絡推定法もいくつか提案されており [10], [11], [12], [13], [14]，その中でも計算コストの少ないものは実時間声質変換技術などに応用されている [15], [16]．

統計的パラメトリック音声合成 [17], [18] や歌声合成においても高精度なスペクトル包絡は重要であり，統計的なアプローチに基づくスペクトル包絡推定法 [19] についても検討がなされている．スペクトル包絡の推定精度は合成される音声の品質にも直結することから，STRAIGHT[10] が様々な研究機関で広く利用されてきた．音声モーフィングなどの声質変換技術 [20] についてもスペクトル包絡が重要な役割を担うことからも，高精度なスペクトル包絡推定に関する研究は現在も重要な価値があるといえる．

本研究では，高品質音声合成方式として提案された TANDEM-STRAIGHT[11] や WORLD[13][*1] で用いられているスペクトル包絡推定法をさらに発展させた方式 CheapTrick[21] の誤差評価を行う．文献 [21] では，F0 の時間変化に対する頑健性についての優位性，および主観評価により分析合成音，F0 制御音のどちらについても提案法が高品質な音声を合成可能であることを示している．近年提案されている非周期性指標推定法 [22] は高精度なスペクトル包絡を要求するため，本稿における評価には，この方法に適したアルゴリズムを選定する狙いもある．誤差評価は，実際の音声には周期音以外の雑音が混入することから SNR に対する頑健性，および CheapTrick は F0 同期窓を用いた波形の切り出しがあることから，F0 推定誤差に対する頑健性を評価し有効性について論じる．

2. スペクトル包絡推定法 CheapTrick の概要

はじめに，本稿で誤差評価を実施するスペクトル包絡推

[1] 山梨大学
, University of Yamanashi, Yamanashi 400–8511, Japan
[a)] mmorise@yamanashi.ac.jp

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

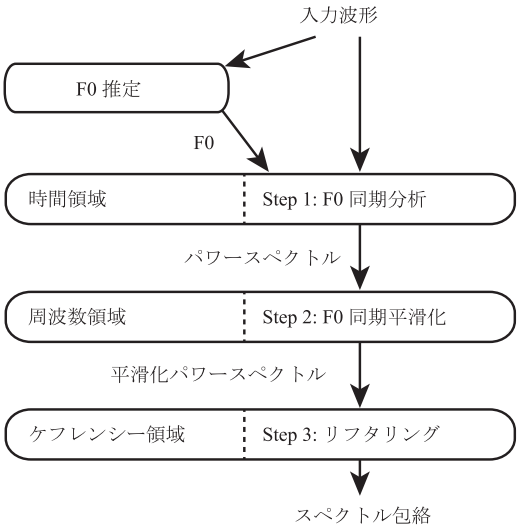

図 1 CheapTrick の処理の手順.

定法 CheapTrick について述べる．高品質音声合成におけるスペクトル包絡推定の問題点は，スペクトル包絡が時間的に変動しない場合でも推定結果が分析時刻に依存して変化することである．STRAIGHT をはじめとするいくつかの方法は，この時間変動成分を定式化し，除去することで高精度な推定を実現してきた．CheapTrick では，この時間変動成分を従来法とは異なる形で定式化し，取り除く方法を提案している．

CheapTrick は，図 1 に示す通り波形と F0 情報を入力とした 3 ステップで構成される．TANDEM-STRAIGHT[11] では，パワースペクトル推定の段階で時間変動成分を取り除き，その後パワースペクトルを適切に平滑化することで時間的に静的かつ滑らかなスペクトル包絡を得ていた．一方，CheapTrick では時間変動成分の除去と平滑化を分離せず，3 ステップを経て最終的に取り除くというコンセプトである．

提案法は，F0 同期分析，F0 同期平滑化，リフタリングの 3 ステップで構成されており，各ステップは最終的な目標である時間的に静的かつ滑らかなスペクトル包絡を得るためのチューニングがなされている．CheapTrick を提案した文献 [21] では，高品質音声合成が目的であることから主観評価を中心にして有効性を論じている．本稿では，CheapTrick の入力となる F0 の推定誤差が結果に与える影響，さらに耐雑音性の観点から客観的な尺度を用いて有効性について検証する．なお，CheapTrick の詳細はオープンアクセスの文献 [21] で示しているため，本稿では概要について述べる．詳細は該当論文を参照されたい．

2.1 ステップ 1: F0 同期分析

CheapTrick では，まずピッチ同期分析 [23] の考えに基づいた窓関数を設計し，切り出された波形のパワースペクトルを計算する．従来法における窓関数の設計とパワースペクトルの計算は，時間的な変動成分を取り除くことを目指して行われてきた．対して，CheapTrick では，パワースペクトル計算の段階で時間変動成分を除去することを目指さず 3 ステップを経て取り除く．ステップ 1 では，基本周期の 3 倍の長さを有する Hanning 窓により波形の切り出しを行いパワースペクトルを計算する．基本周期 (T_0) の 3 倍の Hanning 窓で波形を切り出した場合，以下の式の通り波形のパワーが切り出す時刻に依存せず一定となる．

$$\int_0^{3T_0} (y(t)w(t))^2\,dt \;=\; 1.125\int_0^{T_0} y^2(t)dt. \qquad (1)$$

この Hanning 窓のパワースペクトルは $n\omega_0$ ($\omega_0 = 2\pi/T_0$ であり，n は 0 以外の整数とする) でゼロ点を有する．周期信号のパワースペクトルは $n\omega_0$ Hz でパワー・位相を有する調波構造を有し，窓関数の種類によっては隣接する調波に影響を及ぼす．CheapTrick で用いる窓関数の場合，他の調波に影響を及ぼさない利点もある．ステップ 1 で得られたパワースペクトルは，切り出された波形全体のパワーと，波形の $n\omega_0$ Hz のパワーが時間に対し静的となる．

2.2 ステップ 2: F0 同期平滑化

図 1 より，ステップ 3 ではリフタリング処理，すなわち対数パワースペクトルを用いた処理を行う．ステップ 1 で得られたパワースペクトルにゼロ点が含まれる場合，対数パワースペクトルでは $-\infty$ となりケプストラムの計算，およびリフタリングにおいて悪影響を及ぼす．CheapTrick では，ステップ 2 においてパワースペクトルの平滑化を適切に行うことにより，ステップ 3 で受ける悪影響を低減する．言い換えれば，ステップ 2 は，ステップ 3 の前処理と位置付けられる．

周期信号のパワースペクトルは調波構造を有するため，次数の多いフィルタで平滑化を行うことは調波間の干渉を拡大することとなる．CheapTrick では，最も調波間の干渉を抑えつつ平滑化を行うフィルタとして矩形窓を選択する．具体的には，以下の式により平滑化が実施される．

$$P_s(\omega) \;=\; \frac{3}{2\omega_0}\int_{-\frac{\omega_0}{3}}^{\frac{\omega_0}{3}} P(\omega+\lambda)d\lambda, \qquad (2)$$

ここで，$P(\omega)$ は，ステップ 1 で得られたパワースペクトルを示す．矩形窓の幅は $2\omega_0/3$ であり，これは時間分解能と周波数分解等とのバランスを勘案して経験的に設定した値である．矩形窓の幅が短すぎる場合は，リフタリング処理における悪影響の抑圧が不十分となる．一方，矩形窓の幅を ω_0 にすると時間的な変動を完全に除去することが可能となる一方，周波数分解能が低下しフォルマントが鈍るという別の問題が生じる．

2.3 ステップ 3: リフタリング

ステップ 2 までで得られた平滑化パワースペクトルは，全体のパワーが分析時刻に依存せず，ステップ 2 の平滑化

によりゼロを持たない特徴も有する．一方，時間変動成分は完全に除去されておらず周波数方向の振動成分が残っていることから，ステップ 3 ではこれらの成分を同時に取り除くリフタリング処理を行う．文献 [21] において，分析時刻に依存した成分はケフレンシー領域における nT_0 に集中することが示されているため，リフタリングでは低次の値は保存しつつ nT_0 でゼロを持つ関数で処理することが望まれる．CheapTrick では，この条件を満たす関数として sinc 関数を採用している．ケフレンシー領域において sinc 関数を乗ずることは，対数パワースペクトルを矩形窓を用いて平滑化していることに相当することから，本処理は時間変動成分の除去と平滑化が同時に行われることを意味する．

ステップ 1 の窓関数では各調波のパワーを正確に求めることに成功しているが，ステップ 2 の平滑化，およびステップ 3 における sinc 関数によるリフタリングにより，各調波の振幅には隣接する調波から影響を受けることになる．スペクトル復元は，2 種の平滑化により生じた干渉の影響を取り除くためのリフタリング処理として行われる．ここでは，consistent sampling [24] の理論に基づく方法を導入することとした．従来のサンプリング定理では，サンプリング周波数が fs Hz の場合，入力波形に fs / 2 Hz 以下の成分しか存在しない場合完全に復元可能であることを示している [*2]．対して，consistent sampling の考え方では，サンプリングされたディジタル信号に対し，さらに DA/AD 変換したディジタル信号が同一となることを条件とし，条件を満たすためのフィルタのデザインについて論じている．この条件を完全に満足する場合，無限回 DA/AD 変換しても同一の結果が得られる，すなわちディジタル信号だけではなくアナログ信号も保存可能という利点がある．スペクトル包絡推定にこの考え方を当てはめると，$n\omega_0$ Hz のパワーが AD 変換されたディジタル信号系列であり，スペクトル包絡がアナログ信号系列と位置付けられる．平滑化によりサンプリング点（$n\omega_0$ Hz）の値に誤差が生じるため，それを復元するためのフィルタ処理を行うこととなる．

ある調波のパワーが隣接する調波に与える影響は，1 つ隣に対して Hanning 窓のサイドローブである概ね-30 dB 程度となり，2 つ以上隣に与える影響はさらに指数的に減衰する．実音声には雑音の混入や F0 の揺らぎが生じるため，実用上は隣接する調波の影響のみ除去すれば充分で，それ以上補償した場合は副作用のほうが強くなると判断した．最終的には，以下の式により対数パワースペクトルを $\pm\omega_0$ Hz シフトし，隣接する調波に与えた影響に相当する重み $\tilde{q}_1$ を乗じて加算することで実現した．

$$P_l(\omega) = \exp\left(\tilde{q}_0 \log(P(\omega)) + \tilde{q}_1 \log\left(P(\omega+\omega_0)P(\omega-\omega_0)\right)\right), \tag{3}$$

この操作は，対数パワースペクトルに対し，0 Hz と $\pm\omega_0$ Hz に値を持つデルタ関数を畳み込むことと等しいことから，ケフレンシー領域におけるリフタリング処理に置き換えることが可能になる．

これらの処理を踏まえた最終的なスペクトル包絡 $P_l(\omega)$ は，以下の式により与えられる．

$$P_l(\omega) = \exp\left(\mathcal{F}\left[l_s(\tau)l_q(\tau)p_s(\tau)\right]\right), \tag{4}$$

$$l_s(\tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_0 \tau}, \tag{5}$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_0}\right), \tag{6}$$

$$p_s(\tau) = \mathcal{F}^{-1}\left[\log\left(P_s(\omega)\right)\right], \tag{7}$$

ここで，$l_s(\tau)$ は時間変動を取り除くためのリフタリング関数，$l_q(\tau)$ はスペクトル復元を行うためのリフタリング関数を示す．$\tilde{q}_0$ と $\tilde{q}_1$ はスペクトル復元を行うためのパラメータであり，この係数を調整することにより隣接する調波の影響を除去する．$\mathcal{F}[]$ と $\mathcal{F}^{-1}[]$ は，フーリエ変換とフーリエ逆変換を示す記号である．なお，本稿では，$\tilde{q}_0$ は 1.18，$\tilde{q}_1$ は -0.09 とした．これは予備的に実施した評価から決定した値である．

3. 評価

主観評価はすでに実施しており，CheapTrick は TANDEM-STRAIGHT などの従来法と比較して相対的に高品質な音声が合成可能であることを示している．本稿では，客観的な指標を用いた誤差評価を実施した結果から，CheapTrick の有効性について議論する．具体的には，実音声には非周期的な雑音成分などが混入することから耐雑音性に関する評価，および窓関数設計に必要となる F0 情報の誤差が結果に与える影響について調査することとした．

3.1 評価指標の定義

客観評価は，スペクトル包絡の推定精度に関する指標だけではなく，目的とする分析時刻に依存しない推定が実現できているかも検証するための指標を導入する．はじめに，スペクトル包絡の推定精度に関しては以下の式で求めることとする．

$$E_f = \frac{1}{N}\sum_{n=0}^{N-1}\sigma_f(n), \tag{8}$$

$$\sigma_f^2(n) = \frac{1}{K}\sum_{k=0}^{K-1}\left(P_e(k,n) - \frac{1}{K}\sum_{l=0}^{K-1}P_e(l,n)\right)^2, \tag{9}$$

$$P_e(k,n) = 10\log_{10}\left(P_l(k,n)\right) - 10\log_{10}\left(P_t(k)\right), \tag{10}$$

$P_l(k,n)$ は，各手法で求められたスペクトル包絡の時間周波数表現で，k は離散周波数番号，n はフレーム番号（離散時間）を示す．$P_t(k)$ は真値となるスペクトル包絡に対応し，本評価では時刻に依存せず一定の値とする．また，

[*2] 理想フィルタを仮定するため，現実的には理想通りとならない

K は FFT 長の半分を示し，N は誤差計算に用いるフレーム数を示す．E_f は，各フレームについて真値との対数スペクトル距離を求め，全フレームにおける対数スペクトル距離の平均を計算することとなる．

もう 1 つの評価指標として，分析時刻に対する変動量を以下の式で定義する．

$$E_t = \frac{1}{K}\sum_{k=0}^{K-1}\sigma_t(k), \tag{11}$$

$$\sigma_t^2(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left(P_e(k,n) - \frac{1}{N}\sum_{m=0}^{N-1}P_e(k,m)\right)^2. \tag{12}$$

E_t は，各離散周波数番号について全フレームでの結果の標準偏差を求め，その結果を全周波数について平均する指標となる．本評価では，E_f，E_t のどちらも最小化することが可能な方法が優れたスペクトル包絡推定法と言える．なお，2 つの指標の片方の誤差が小さく片方が大きくなることもあり得るが，品質に与える影響については調査せず，2 つの個別な指標で評価した結果に基づいて論じる．

3.2 実験条件

CheapTrick と比較する方法として，高品質音声合成方式である TANDEM-STRAIGHT と，筆者らが提案した CheapTrick の前身となる STAR [13] を採用した．STAR は，CheapTrick におけるステップ 2 の処理，およびステップ 3 におけるスペクトル復元を行わない以外，CheapTrick と同一となる．前報において合成音声の品質に関する評価，すなわち時変的なスペクトル包絡に対する有効性は示されているため，本稿では，耐雑音性と F0 の推定誤差に特化して誤差評価を実施する．分析対象とする信号は，特定の基本周波数 (Standard F0) を有するパルス列とする．すなわち，$P_t(k)$ はフラットとなるため，全時刻についてフラットな特性を得ることが求められる．信号のサンプリング周波数は 48 kHz，信号長は 1 s とした．分析シフト幅は 1 ms としたことから，フレーム数 N は 1,000 となる．また，FFT 長は，TANDEM-STRAIGHT などがサポートする F0 の下限が 40 Hz であり，その 3 倍の窓長が必要となる条件から 4,096，K は 2,048 に設定された．なお，F0 の違いが結果に与える影響を確認するため，Standard F0 は 100 Hz と 200 Hz の 2 種類とした．

3.3 実験 1: 耐雑音性に関する評価

耐雑音性に関する評価では，SNR が 0 dB から 60 dB まで 0.1 dB 刻みに設定し，加算する雑音はホワイトノイズとした．また，1 条件について 100 種類のホワイトノイズを用いて計算し，最終的な結果は 100 回の平均値とした．スペクトル包絡の真値がフラットであるため，ホワイトノイズを用いることで全帯域における SNR は概ね一致することとなる．図 2，3 に評価結果を示す．図の横軸は SNR を示し，縦軸は各評価指標を示す．具体的に，上図は推定精度に関する評価指標 E_f を示し，下図は時間変動に関する評価指標 E_t を示す．

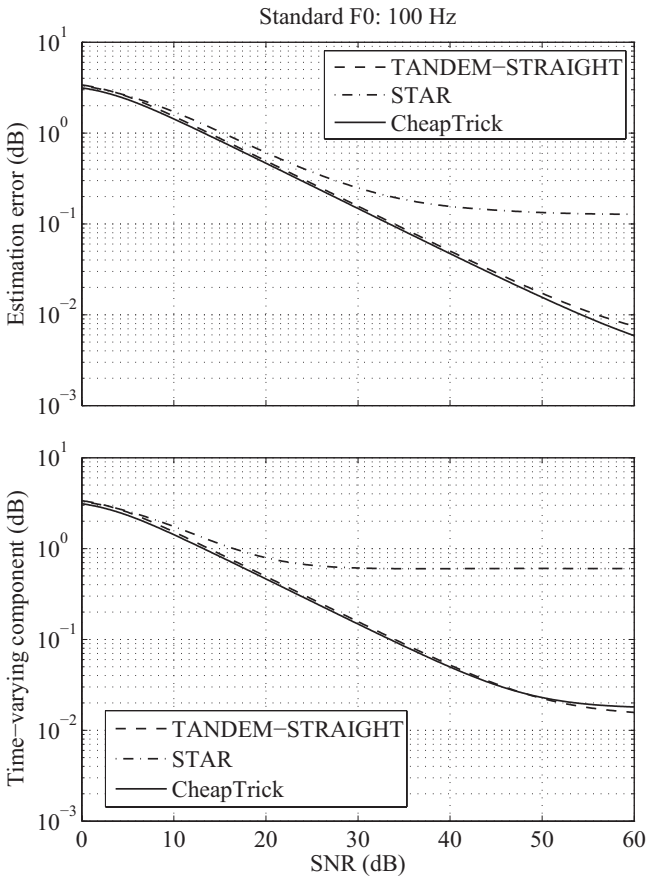

図 2 SNR と推定性能との関係．Standard F0 が 100 Hz の場合の結果．

推定精度に関する評価指標では，SNR，Standard F0 に関わらず CheapTrick，TANDEM-STRAIGHT，STAR の順に性能が高いことを示す．一方，時間変動に関する評価指標では，SNR が 45 dB 以上になると TANDEM-STRAIGHT が最も高い性能を示すことが確認できる．

3.4 実験 2: F0 の誤差に対する頑健性

F0 の誤差が推定結果に与える提供を調査するため，−20 から 20%の推定誤差について 0.1%刻みで各誤差を計算した．また，現実的な条件で評価するため，本評価における SNR を 60 dB とした．図 4，5 に実験結果を示す．推定精度に関する結果から，STAR が他の 2 手法よりも低い性能であることが示される一方，CheapTrick と TANDEM-STRAIGHT はほぼ等価な性能であるといえる．時間変動に関しては，F0 の誤差が ±3%以上の場合 CheapTrick が明らかに高い性能を示すことが確認できる．また，TANDEM-STRAIGHT の結果に関して，時間変動の評価指標が F0 誤差に対して階段状に変化していることも確認できる．この原因については，TANDEM-STRAIGHT で窓関数を設計

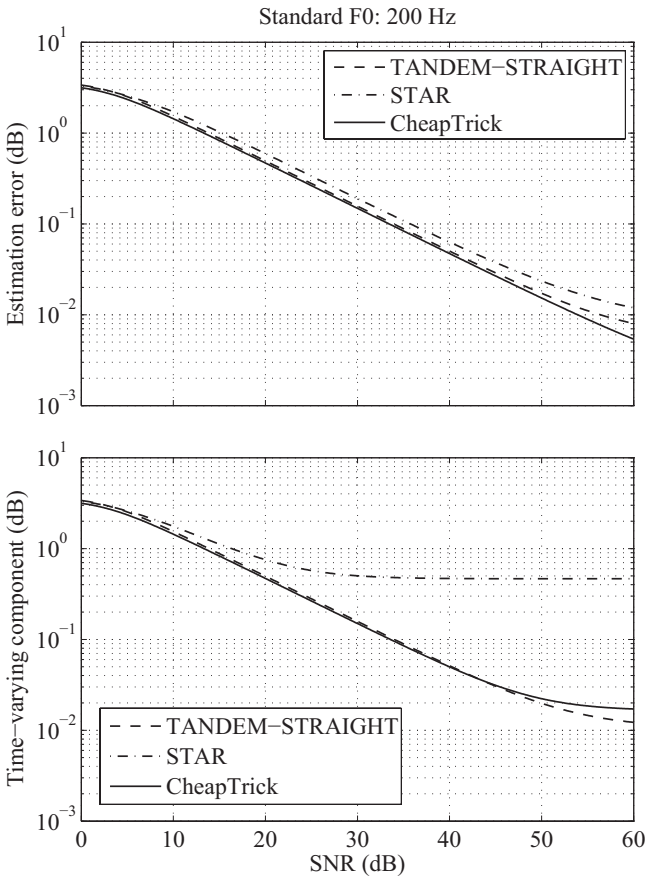

図 3 SNR と推定性能との関係．Standard F0 が 200 Hz の場合の結果.

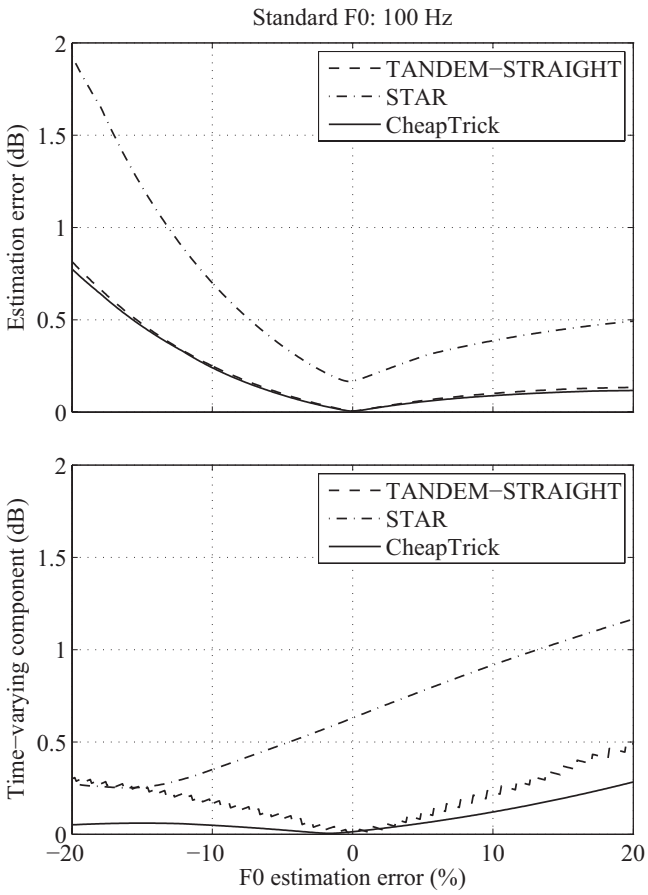

図 4 F0 の推定誤差が推定結果に与える影響．Standard F0 が 100 Hz の場合の結果.

する際，窓長が変化したタイミングで階段状の変化が生じていることであった [*3].

3.5 考察

これらの実験結果から，CheapTrick は，SNR が 45 dB 以上の場合を除き TANDEM-STRAIGHT と等価以上の性能を達成することが示された．すでに実施されている主観評価の結果，CheapTrick は最も高品質な音声を合成可能であることから，実音声には F0 の揺らぎや雑音の影響が SNR 45 dB よりも大きく，実用上は CheapTrick で十分といえる．特に，TANDEM-STRAIGHT と顕著な差が生じたのは時間変動の評価指標であることから，高品質な音声合成を目的としたスペクトル包絡推定において，時間的に静的なスペクトル包絡を推定することが重要であることを示唆する．今後は，CheapTrick を用いた統計的パラメトリック歌声・音声合成や，音声特徴量抽出への応用について検討する必要があるといえる．

4. おわりに

本稿では，筆者が提案したスペクトル包絡推定法 CheapTrick について，入力パラメータである F0 の誤差と加算性雑音に対する頑健性を評価した．はじめに，CheapTrick の概要を説明し，その後 2 つの誤差評価により提案法の有効性について検証した．結果から，提案法は，雑音に対する頑健性について実用的な条件において従来法よりも優れていることを示した．また，F0 の誤差に対する頑健性の評価では，スペクトル包絡の推定性能は TANDEM-STRAIGHT とほぼ等価であるが，分析時刻に対する依存性では F0 の誤差に対する影響が相対的に小さいことを示した．以上の結果から，提案法は，高品質な音声や歌声を合成できるだけではなく，音声パラメータ推定の客観的な性能においても良好であることが示された．

今後の検討課題には，CheapTrick を用いた声質変換技術や統計的パラメトリック合成への適用が挙げられる．統計的パラメトリック音声合成においてスペクトル包絡推定の貢献は大きいため，STRAIGHT に代わるスペクトル包絡推定法として充分な品質を達成可能か検証する見通しである．また，我々が現在検討して時間的に静的な群遅延推定においてもスペクトル包絡推定法が必要となるため [22]，静的群遅延推定の精度の視点からも CheapTrick の有効性を検証する．

謝辞 本研究の一部は，JSPS 科研費 24300073,

[*3] TANDEM-STRAIGHT では，同一の窓関数長でも F0 に基づいた微調整を行って窓関数を設計していること原因であるが，本評価でその修正は行っていない.

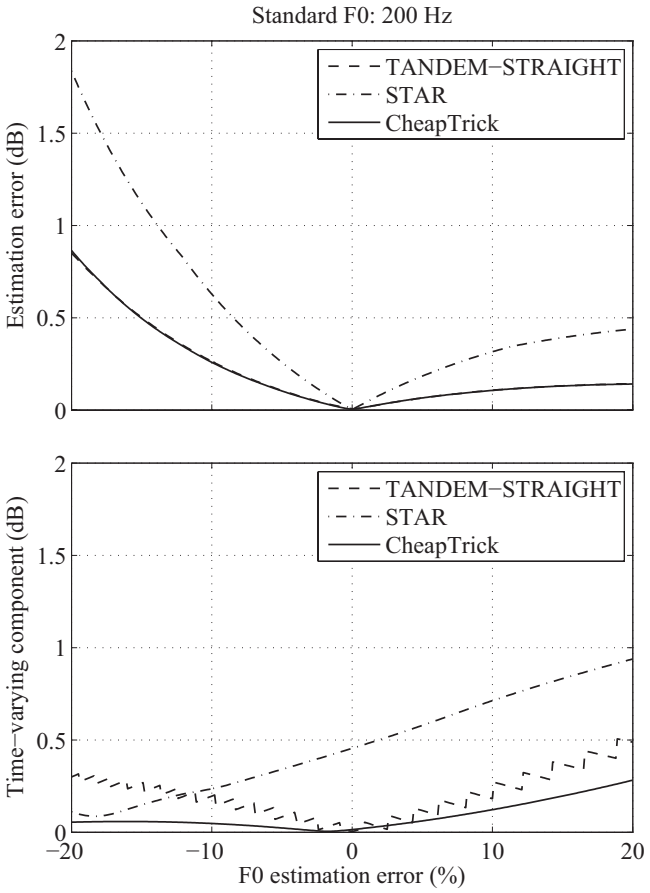

図 5 F0 の推定誤差が推定結果に与える影響．Standard F0 が 200 Hz の場合の結果．

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）の支援を受けた．

参考文献

[1] Moulines, E. and Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication*, Vol. 9, No. 5-6, pp. 453–467 (1990).

[2] McAulay, R. and Quatieri, T.: Speech analysis/synthesis based on a sinusoidal representation, *IEEE Trans. Acoustics, Speech, and Signal Processing)*, Vol. 34, No. 4, pp. 744–755 (1986).

[3] Dudley, H.: Remaking speech, *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177 (1939).

[4] Noll, A. M.: Short-time spectrum and “cepstrum” techniques for vocal pitch detection, *J. Acoust. Soc. Am.*, Vol. 36, No. 2, pp. 296–302 (1964).

[5] Oppenheim, A. V.: Speech analysis-synthesis system based on homomorphic filtering, *J. Acoust. Soc. Am.*, Vol. 45, No. 2, pp. 458–465 (1969).

[6] Atal, B. S. and Hanauer, S. L.: Speech analysis and synthesis by linear prediction of the speech wave, *J. Acoust. Soc. Am.*, Vol. 50, No. 2B, pp. 637–655 (1969).

[7] El-Jaroudi, A. and Makhoul, J.: Discrete all-pole modeling, *IEEE Trans. on Signal Processing*, Vol. 39, No. 2, pp. 411–423 (1991).

[8] Badeau, R. and David, B.: Weighted maximum likelihood autoregressive and moving average spectrum modeling, *in Proc. ICASSP 2008*, pp. 3761–3764 (2009).

[9] Campedel-Oudot, M., Cappé, O. and Moulines, E.: Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach, *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 5, pp. 469–481 (2001).

[10] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,” Speech Communication, *Speech Communication*, Vol. 27, pp. 187–207 (1999).

[11] 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. ICASSP 2008*, pp. 3933–3936 (2008).

[12] 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, pp. 713–728 (2011).

[13] Morise, M.: An attempt to develop a singing synthesizer by collaborative creation, *in Proc. SMAC 2013*, pp. 287–292 (2013).

[14] 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).

[15] Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T. and Kawahara, H.: Implementation of realtime STRAIGHT speech manipulation system, *Acoust. Sci. & Tech.*, Vol. 28, No. 3, pp. 140–146 (2007).

[16] Morise, M., Onishi, M., Kawahara, H. and Katayose, H.: v.morish’09: A morphing-based singing design interface for vocal melodies, *Lecture Notes in Computer Science*, Vol. LNCS 5709, pp. 185–190 (2009).

[17] Zen, H., Tokuda, K. and Black, A. W.: Statistical parametric speech synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).

[18] Zen, H., Senior, A. and Schuster, M.: Statistical parametric speech synthesis using deep neural networks, *in Proc. ICASSP 2013*, pp. 7962–7966 (2013).

[19] Toda, T. and Tokuda, K.: Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM, *in Proc. ICASSP 2008*, pp. 3925–3928 (2008).

[20] Kawahara, H., Nisimura, R., Irino, T., Morise, M., Takahashi, T. and Banno, H.: Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown, *in Proc. ICASSP2009*, pp. 3905–3908 (2009).

[21] Morise, M.: CheapTrick, a spectral envelope estimator for high-quality speech synthesis, *Speech Communication* (2014).

[22] Kawahara, H., Morise, M., Banno, H., Nisimura, R. and Irino, T.: Excitation source analysis for high-quality speech manipulation systems based on an interference-free representation of group delay with minimum phase response compensation, pp. xxx–xxx (2014).

[23] Mathews, M. V., Miller, J. E. and David, E. E.: Pitch synchronous analysis of voiced sounds, *J. Acoust. Soc. Am.*, Vol. 33, pp. 179–186 (1961).

[24] Unser, M.: Sampling — 50 years after Shannon, *Proc. of the IEEE*, Vol. 88, pp. 569–587 (2000).