高速近似連続ウェーブレット変換による振幅スペクトログラムに対する実時間位相推定法*

◎中村友彦（東大院・情報理工），亀岡弘和（東大院・情報理工，NTT・CS 研）

1 はじめに

定 Q フィルタバンクとしても知られる連続ウェーブレット変換（CWT）は，対数周波数スケールで一様な解像度をもつ時間周波数表現を与える．人間の聴覚システムも特に高周波数帯域で対数的な周波数解像度を持つことが知られており，聴覚システムに着想を得た音響信号処理手法の開発には CWT で得られたスペクトログラム上での処理が望ましいと考えられる．実際に，複素スペクトログラムを用いたマルチチャネル音源分離 [1]，振幅スペクトログラムを用いた多重基本周波数推定 [2–4] や歌声分離 [5] などで CWT の有用性が確認されている．音源分離や音響信号加工などの時間領域の音響信号を生成することが目的のアプリケーションにおいては，分離や加工された振幅スペクトログラムを音響信号に変換する必要がある．そのため，与えられた振幅スペクトログラムに対し適切な位相を与える必要がある．

最初に提案された振幅スペクトログラムからの位相推定アルゴリズム [6] は，所望の振幅スペクトログラムに現在の位相の推定値を割り当てた複素スペクトログラムに逆 CWT と CWT を行うステップと，複素スペクトログラムの振幅部分のみを所望の振幅スペクトログラムと置換するステップからなり，これらのステップを繰り返し行うことで位相を推定する．しかし，CWT は計算量が高いためこのアルゴリズムは長時間の計算を要し [7]，実際のアプリケーションに用いるためには計算量の削減が重要となる．また，アルゴリズムの収束性も実用上重要である．CWT と逆 CWT の様々な高速計算法は近年相次いで提案されており [8–10]，CWT の代わりにこれらの高速計算法を利用することで計算法を削減できるはずである．しかし，そのようなアルゴリズムでは収束性が保証されるかどうかが不明であった．

そこで，我々は補助関数法 [11] と呼ばれる最適化原理に基づき導出したアルゴリズムが [6] で提案されたアルゴリズムと一致することを示し，冗長な線形変換であればアルゴリズムの収束性が保証されることを明らかにした [7]．この結果を元に，CWT の代わりに高速近似 CWT [8] を用いることで，収束性を保証しつつ計算量の削減されたアルゴリズム（高速位相推定アルゴリズム）を提案した．

高速近似 CWT を含め多くの CWT 高速化アルゴリズムでは信号全体の高速 Fourier 変換（FFT）を用いることが前提となるため，信号長が長いほど高速位相推定アルゴリズムの空間計算量は増加する．そのため、実際的な信号長のスペクトログラムを計算するためには多くのメモリが必要となり，オーディオプレーヤーなどのメモリ制限のある装置では処理可能な音響信号が限られる．また，原理上実時間動作を要求するアプリケーションにそのまま適用することはできない．

そこで本稿では，高速位相推定アルゴリズムを拡張し，信号長によらず一定の空間計算量をもち音響信号を逐次処理可能な実時間位相推定アルゴリズムを提案する．短時間 Fourier 変換で得られた振幅スペクトログラムからの実時間位相推定アルゴリズム [12] と同様に，提案アルゴリズムでは音響信号をオーバーラップのある固定長のフレームに分割し，各フレームに高速近似 CWT を適用して得られたスペクトログラムを対象とする．高速位相推定アルゴリズムは位相全体へ定数を加えたものも解として許容しうるため，隣り合うフレーム同士で独立に高速位相推定アルゴリズムを適用した場合には少なくとも異なる初期位相をとりうる．そのため，フレーム同士のオーバーラップ部分で不連続な信号成分となる位相が推定される可能性がある．そこで，提案アルゴリズムでは隣り合うフレーム間のオーバーラップ部分の信号成分が同一である（無矛盾）ことを考慮することにより，不連続な信号となることを抑制する．

2 高速近似ウェーブレット変換を用いた位相推定アルゴリズム

2.1 高速近似 CWT

まず，高速近似 CWT について簡単に紹介する．CWT は，スケーリングされたアナライジングウェーブレットが各フィルタのインパルス応答に対応するフィルタバンクと解釈できる．Morlet や対数正規分布型ウェーブレット [3] などのアナライジングウェーブレットを用いた場合，各サブバンドフィルタの周波数特性の主要な部分が局所的に分布する．そのため，主要な値が存在する周波数領域のみを計算に用いることで時間信号の CWT を高速に計算できる．高速近似 CWT はこの考えに基づき提案された．

T 次元の時間信号を $\boldsymbol{f}$，T 次元の DFT 行列を F_T とする．高速近似 CWT では，時間信号全体の FFT，$F_T\boldsymbol{f}$ を求めた後に各サブバンドフィルタの主要な部分が存在する領域 $k \in [B, B+D-1]$ に帯域制限を行う．ここで，$k = 0, \cdots, T-1$ は角周波数インデックスである．ある対数周波数インデックス $\omega(= 0, \cdots, \Omega-1)$ に対応するサブバンドフィルタについて考える．帯域制限を表

---

* Real-Time Phase Estimation from Magnitude Spectrogram of Fast Approximate Continuous Wavelet Transform by NAKAMURA, Tomohiko (The University of Tokyo) and KAMEOKA, Hirokazu (The University of Tokyo / Nippon Telegraph and Telephone Corporation)

す行列は，全要素が 0 の $D\times B$ 次元の行列 $0_{D\times B}$ と D 次元の単位行列 I_D を用いて，$L := [0_{D\times B}, I_D, 0_{D\times(T-D-B)}]$ と書ける．帯域制限された時間信号の FFT に対して，帯域制限された周波数帯域の当該フィルタの周波数特性 $\boldsymbol{\psi}_\omega \in \mathbb{C}^D$ を乗算し，$\mathrm{diag}(\boldsymbol{\psi}_\omega)LF_T\boldsymbol{f}$ を得る．ここで，$\mathrm{diag}(\boldsymbol{\psi}_\omega)$ は $\boldsymbol{\psi}_\omega$ を対角成分に並べた対角行列を表す．帯域制限により正規化角周波数が $[2\pi B/D, 2\pi B/D+2\pi]$ に分布するため，$\mathrm{diag}(\boldsymbol{\psi}_\omega)LF_T\boldsymbol{f}$ を巡回的にシフトさせ，先頭の成分の位相が 0 となるようにする．この操作は以下の行列 C で書ける．

$$C := \begin{bmatrix} 0_{(B-(p-1)D)\times(pD-B)} & I_{B-(p-1)D} \\ I_{(pD-B)} & 0_{(pD-B)\times(B-(p-1)D)} \end{bmatrix} \quad (1)$$

ここで，p は $pD \le B + D < (p+1)D$ となる最大の整数である．巡回シフトを行ったものに対し D 次元の逆 FFT を行うことで，当該サブバンドでのスペクトログラムが得られる．

これらをまとめると，対数周波数インデックス ω に対応する高速近似 CWT $W_\omega = F_D^{\mathsf{H}} C\,\mathrm{diag}(\boldsymbol{\psi}_\omega)L$ を用いて，高速近似 CWT は $W := [W_0^\top, \cdots, W_{\Omega-1}]^\top$ と表せる．ここで，$^{\mathsf{H}}$ は当該変数の Hermitian 共役を表す．この逆変換（逆高速近似 CWT）は W の擬似逆行列 W^+ で表される．D, B はサブバンド毎に異なってもよいが，簡単のため以下では全てのサブバンドで D, B は同一とする．

2.2 高速位相推定アルゴリズム

W は T 次元複素ベクトルから $D\Omega$ 次元複素ベクトルへの変換であるため，$T < D\Omega$ の場合，高速近似 CWT で得られるスペクトログラムは時間信号の冗長な表現である．したがって，スペクトログラムの集合 $\mathcal{W}$ は $\mathbb{C}^{D\Omega}$ の部分空間である．ここで，逆高速近似 CWT の後高速近似 CWT を適用する操作 WW^+ を考える．WW^+ は $\mathcal{W}$ への直行射影であるため，$D\Omega$ 次元の複素ベクトル $\boldsymbol{y}$ に対し $\boldsymbol{y} - WW^+\boldsymbol{y}$ は $\boldsymbol{y}$ から $\mathcal{W}$ への距離を表す．したがって，この距離が小さければ小さいほど $\boldsymbol{y}$ はより「スペクトログラムらしい」とみなせ，$\boldsymbol{y} = WW^+\boldsymbol{y}$ のとき $\boldsymbol{y}$ はスペクトログラムであり，対応する時間信号が存在する．そのため，与えられた振幅スペクトログラムに対して，この距離を最小化する位相を求める問題として位相推定問題は定式化できる．

振幅スペクトログラム $\boldsymbol{a} \in \mathbb{R}_{\ge 0}^{\Omega T}$ が与えられたとき，位相推定問題は位相の推定値 $\boldsymbol{\phi} \in [0, 2\pi)^{\Omega T}$ を用いて以下のように定式化できる．

$$\min_{\boldsymbol{\phi}\in[0,2\pi)^{\Omega T}} \boldsymbol{s}^{\mathsf{H}}(\boldsymbol{a},\boldsymbol{\phi})(I_{\Omega T} - WW^+)\boldsymbol{s}(\boldsymbol{a},\boldsymbol{\phi}) \quad (2)$$

ここで，$\boldsymbol{s}(\boldsymbol{a},\boldsymbol{\phi})$ は振幅 $\boldsymbol{a}$，位相 $\boldsymbol{\phi}$ の複素ベクトルを表す．この問題は $\boldsymbol{s}(\boldsymbol{a},\boldsymbol{\phi})$ 自体をパラメータ $\boldsymbol{s} = [s_{0,0}, s_{0,1}, \cdots, s_{0,D}, s_{1,0}, \cdots, s_{\Omega-1,D}]^\top$ とみなせば，各時間周波数成分 $s_{\omega,t}$ に関し $|s_{\omega,t}|^2 = a_{\omega,t}^2$ という 2 次制約を持つ 2 次計画問題に書き換えられる．しかし，WW^+ の行数および列数はスペクトログラムの要素数と同一であるため計算コストが非常に高く，実際的にはこの 2 次計画問題を直接解くことは難しい．

そこで，我々は [7] で補助関数法と呼ばれる最適化原理に基づき高速位相推定アルゴリズムを提案した．導出の詳細は省くが，下記の 2 つの更新式を交互に更新することにより，広義単調減少させることができる．

$$\tilde{\boldsymbol{s}} \leftarrow WW^+\boldsymbol{s}(\boldsymbol{a},\boldsymbol{\phi}) \quad (3)$$

$$\boldsymbol{\phi} \leftarrow \angle\tilde{\boldsymbol{s}} \quad (4)$$

ここで，$\angle\tilde{\boldsymbol{s}}$ は複素ベクトル $\tilde{\boldsymbol{s}}$ の各要素の偏角を $[0, 2\pi)^{\Omega T}$ のベクトルとして返す演算子である．(3) 式は，現在のスペクトログラム推定値に対し逆高速近似 CWT を行ったのちに高速近似 CWT を適用することを表す．(4) 式は，(3) 式の操作で得られたスペクトログラム $\tilde{\boldsymbol{s}}$ の位相を新たな位相の推定値とすることに対応する．

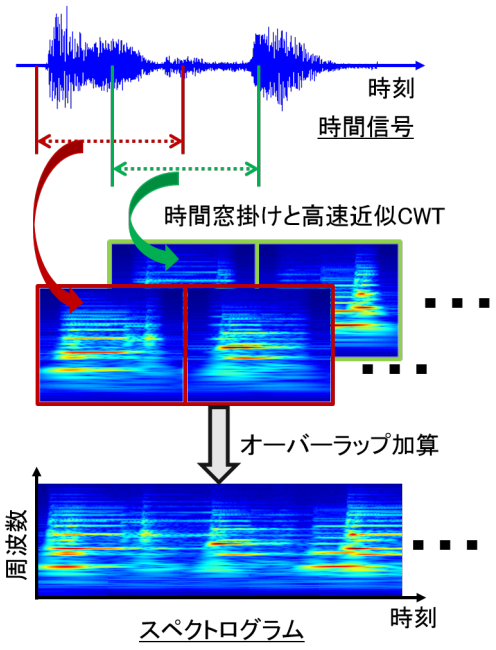


Fig. 1: オンライン高速近似 CWT の処理フロー．

3 位相推定アルゴリズムの実時間化

3.1 オンライン高速近似 CWT

前節までの処理では高速近似 CWT に時間信号全体の FFT が必要であるため，音響信号が逐次入力される場合には適用できない．このように場合には逐次処理が可能な CWT の高速計算法 [9] を用いることができ，本節では [9] と同一の方法で高速近似 CWT を拡張する．この手法をオンライン高速近似 CWT と呼ぶ．

オンライン高速近似 CWT は 3 ステップの処理からなる（図 1）．(i) フレーム長 $2N$，フレームシフト N として，入力音響信号の当該時刻でのフレームを切り出す．(ii) 当該フレームの信号成分に分析窓 $\boldsymbol{h} = [h_0, \cdots, h_{2N-1}]^\top$ を掛けたものに対して，高速近

似 CWT を適用して当該フレームでのスペクトログラムを得る．(iii) 得られた各フレームでのスペクトログラム同士を対応する時刻で重畳することにより所望のスペクトログラムが得られる．

逆変換では，ステップ (ii) で得られた当該フレームのスペクトログラムと，当該フレームと直前のフレームのオーバーラップ部分の時間信号があればよい．当該フレームのスペクトログラムに逆高速近似 CWT を適用し得られた時間信号に合成窓 $\boldsymbol{v} = [v_0, \cdots, v_{2N-1}]^\top$ を掛け，適切にオーバーラップ加算することで入力された時間信号が得られる．ここで，合成窓は時刻インデックス $n = 0, \cdots, N-1$ において $h_n v_n + h_{n+N} v_{n+N} = 1$ を満たす．そのため，空間計算量の主な増大要因である各フレームのスペクトログラムの要素数は信号長に依存せず，信号長に比べ小さな N を用いれば空間計算量が削減できる．

ここで，h_n として

$$h_n = \begin{cases} 0 & (0 \le n < \frac{N-M}{2}) \\ 0.5 - 0.5\cos\left(\pi \dfrac{n - \frac{N-M}{2}}{M-1}\right) & (\frac{N-M}{2} \le n < \frac{N+M}{2}) \\ 1 & (\frac{N+M}{2} \le n < \frac{3N-M}{2}) \\ 0.5 + 0.5\cos\left(\pi \dfrac{n - \frac{3N-M}{2}}{M-1}\right) & (\frac{3N-M}{2} \le n < \frac{3N+M}{2}) \\ 0 & (\frac{3N+M}{2} \le n < 2N) \end{cases} \tag{5}$$

で定義される Tukey 窓を用いると，$n = 0, 1, \cdots, N-1$ において $h_n + h_{n+N} = 1$ が成り立つため各 $n = 0, \cdots, 2N-1$ で $v_n = 1$ となる．M $(0 < M < N)$ はオーバーラップ量を調節するパラメータであり，M が大きいほどフレーム同士のオーバーラップ量が多い．

3.2 実時間位相推定アルゴリズム

高速位相推定アルゴリズムは (2) 式から分かる通り位相全体へ定数を加えたものも解として許容するため，隣り合うフレーム同士で独立に高速位相推定アルゴリズムを適用した場合には少なくとも異なる初期位相をとりうる．そのため，フレーム毎に高速位相推定アルゴリズムを独立に適用すると，フレーム同士のオーバーラップ部分で不連続な信号成分となる位相が推定される可能性が高い．このような信号成分の不連続性を抑制するために，隣接フレーム間のオーバーラップ部分の信号成分が無矛盾であることを考慮したアルゴリズムを提案する．

直前のフレームで得られた時間信号の当該フレームに対応する部分を，N 次元の複素ベクトル $\boldsymbol{g}$ で表す．$\boldsymbol{g}$ のうち直前のフレーム内の時刻に含まれない当該フレームの部分には 0 を詰める．$\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ を当該フレームのスペクトログラムの推定値として再定義すると，$\boldsymbol{g}$ が当該フレームの時間信号と無矛盾であれば，

$$W^+ \boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) = \mathrm{diag}(\boldsymbol{h})\mathrm{diag}(\boldsymbol{v})(\boldsymbol{g} + W^+ \boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})) \tag{6}$$

が成り立つ．そこで，(6) 式の右辺を $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ に対する時間信号への逆変換とみなし，(3) 式の代わりに

$$\tilde{\boldsymbol{s}} \leftarrow W\mathrm{diag}(\boldsymbol{h})\mathrm{diag}(\boldsymbol{v})(\boldsymbol{g} + W^+ \boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})) \tag{7}$$

を用いることで，フレーム同士のオーバーラップ部分で不連続な信号成分となることを抑制できると考えられる．本稿では，このアルゴリズムを実時間位相推定アルゴリズムと呼ぶ．

このアルゴリズムでは，当該フレームのスペクトログラムと $\boldsymbol{g}$ のみを保存すればよく，空間計算量は $O(N + \Omega D)$ である．そのため，空間計算量が信号長に依存せず，信号長が大きい場合に高速位相推定アルゴリズムに比べて空間計算量を格段に削減できる．

4 信号再構成実験による性能評価

4.1 実験条件

フレーム間の無矛盾性の考慮による効果と計算速度を評価するため，振幅スペクトログラムからの時間信号の再構成実験を行った．比較手法として，各フレームに対してそれぞれ高速位相推定アルゴリズムを独立に適用する方法（ベースライン法）を用いた．実験データとして，RWC 音楽ジャンルデータベース [13] の 10 曲を用いた（サンプリング周波数 48 kHz）．各曲の冒頭 30 s にオンライン高速 CWT を適用し得られた振幅スペクトログラムを入力とし，位相はランダムに初期化した．フレーム長は 170, 340, 680 ms（$N = 2^{12}, 2^{13}, 2^{14}$）とし，(5) 式で定義される分析窓を用いた（$M = N/4$）．アナライジングウェーブレットとして対数正規分布型のウェーブレット [3] を用いた．このウェーブレットの Fourier 変換は対数周波数領域で正規分布と同形であり，正規分布の標準偏差に対応するパラメータを 0.02 とした．フィルタの中心周波数が 25 cent 毎に 27.5 から 23679.5 Hz となるようスケールを設計し，高速近似 CWT での帯域制限の範囲は中心周波数から対数周波数上で $[-2\sigma, 2\sigma]$ とした．

計算速度の評価指標として，フレームシフトに対する処理時間の比で定義される real time factor（RTF）を用いた．RTF が 1 以下であれば実時間で実行可能であり，低ければ低いほど高速である．再構成信号の音質の評価指標として，perceptual evaluation of audio quality（PEAQ）[14] による objective difference grade（ODG）を用いた．ODG は −4 から 0 までの値をとり，ODG が大きいほど音質が高い．

4.2 結果

図 2 に，提案法とベースライン法による再構成信号の ODG と計算時間の結果を示す．提案法はベースライン法に比べ高い ODG で時間信号を再構成できており，フレーム間の無矛盾性が再構成信号の音質に重要であることが確認できる．実際に筆者がベースライン法の再構成信号を聴取したところ，フレームのオーバーラップ部分で音量が下がったように聴こえた．これは，オーバーラップ部分で互いに打ち消し合うような信号成分が得られたことが原因であると考

えられる．短いフレーム長ほどベースライン法で得られた再構成信号の ODG が低い傾向があるが，これはフレーム長が短くなるほどフレーム数が増えオーバーラップする信号成分が増えるためにフレーム間の無矛盾性が増しやすいからである．また，フレーム長 340, 680 ms とした提案法では実時間で ODG が −2 以上の再構成信号が得られており，実時間制約下でも実用上十分な音質の音響信号が再構成可能であった．提案アルゴリズムの収束性は理論的には保証されていないが，実験において各フレームでの目的関数は単調に減少したことを確認しており，収束性が保証される可能性が示唆された．

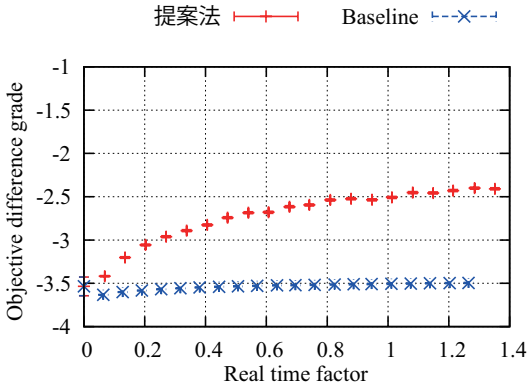
5 まとめ

本稿では振幅スペクトログラムからの実時間位相推定アルゴリズムを提案した．提案アルゴリズムでは固定長のフレーム毎に位相推定を行うことにより，空間計算量が信号長に依存せず長い信号に対しても位相を推定できる．隣接フレーム間でのオーバーラップ部分の信号成分に対する無矛盾性を考慮した更新式を用いることにより，各フレーム対して独立に高速位相推定を適用した場合よりも高音質な信号を振幅スペクトログラムから再構成できることを実験により確認した．また，音楽音響信号に対して実時間制約下でも実用上十分な音質の音響信号が提案アルゴリズムにより再構成できることも確認した．今後は，音声に関する性能評価実験や提案アルゴリズムの収束性に関する理論的な保証が課題である．

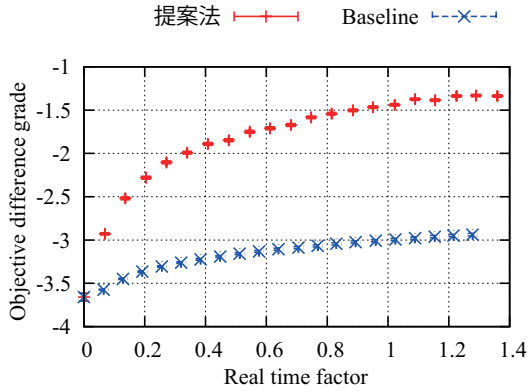
謝辞　本研究は JSPS 科研費 26730100，15J0992 の助成を受けたものです．

参考文献

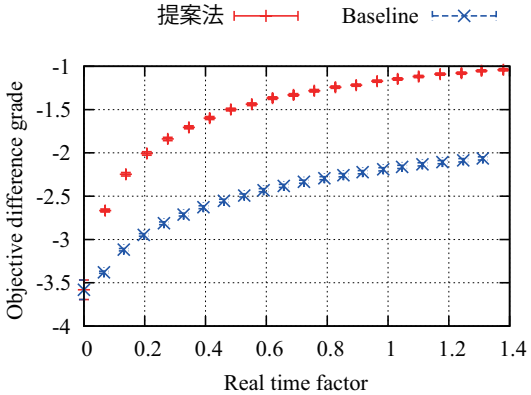
[1] J. J. Burred *et al.*, *Proc. AES Convension*, 2006.
[2] M. N. Schmidt *et al.*, *Proc. ICA-BSS*, pp. 700–707, 2006.
[3] H. Kameoka, Ph.D. dissertation, The University of Tokyo, Mar. 2007.
[4] J. P. de León *et al.*, *Proc. DAFx*, pp. 47–54, 2013.
[5] Y. Ikemiya *et al.*, *Proc. ICASSP*, pp. 574–578, 2015.
[6] T. Irino *et al.*, *IEEE Trans. SP*, 41 (12), pp. 3549–3554, 1993.
[7] T. Nakamura *et al.*, *Proc. DAFx*, pp. 129–135, 2014.
[8] 亀岡他, “信号処理方法及び装置,” Nov. 2008, 特特開 2008-281898.
[9] N. Holighaus *et al.*, *IEEE Trans. ASLP*, 21 (4), pp. 775–785, Apr. 2013.
[10] C. Schörkhuber *et al.*, *Proc. AES Int. Conf. Semantic Audio*, Jan. 2014.
[11] J. M. Ortega *et al.*, *Iterative solution of nonlinear equations in several variables*, no. 30, 1970.
[12] J. Le Roux *et al.*, *Proc. ISCA Tutorial and Research Workshops on Statistical and Perceptual Audition*, pp. 23–28, 2008.
[13] M. Goto, *Proc. Int. Congress Acoust.*, 1, pp. 553–556, 2004.
[14] “ITU-T recommendation BS.1387-1, Perceptual evaluation of audio quality (PEAQ): Method for objective measurements of perceived audio quality,” Sep. 2001.



(a) フレーム長 170 ms



(b) フレーム長 340 ms



(c) フレーム長 680 ms

Fig. 2: 提案法とベースライン法（Baseline）による様々なフレーム長での RTF に対する ODG の平均値と標準誤差．各点は，RTF が小さい方からそれぞれ各フレームで $0, 10, \cdots, 200$ 反復だけアルゴリズムを実行した場合の結果である．