2-2 基本周波数推定（歌声研究に関する視点から）

（執筆者：森勢 将雅）

ピッチは音の高さに対応する心理量であり，ピッチに相当する物理量の基本周波数 (f_0) の推定は，楽器音や音声の音色に相当するスペクトル包絡推定と並び，古くから研究されてきた．基本周波数の抽出技術には，自動採譜，歌声の分析[1)]や合成[2)]，演奏の表情の分析，あるいは，音響に反応するインタラクティブシステムの制作等、幅広い応用領域がある．

基本周波数の推定は，単音楽器や音声単独を対象とした研究と，楽曲中の歌唱など多重音を対象とした研究により方針が大幅に異なる．本稿では，基本周波数推定に関して積極的な研究がなされてきた音声領域の研究事例に焦点を当てて技術紹介を行う．

2-2-1 音声に対する基本周波数の定義と推定に対する問題点

音声は，本来声帯振動を伴う有声音と伴わない無声音とに区別されるが，本稿では有声音に限定して議論する．基本周波数は，声帯振動が生じる時間間隔の逆数と定義され，基本周波数の高低はピッチの高低と対応する．一般的に，人間の発話を長期的に観測すると特性が大きく変化するため，厳密な意味での基本周波数は定義が困難である．通常の基本周波数推定は，音声波形を短時間のフレームとして切り出し，その区間に存在する周期を推定する．しかしながら，短時間の音声を観測した場合も声帯振動の時間間隔や声帯振動の波形は微細に変化しているため，正確な基本周波数の推定は容易ではない．

基本周波数推定法では，音声の時間波形に対する周期性に着目した分析法と，パワースペクトルの調波構造に着目した方法とに大別される[3)]．基本周波数分析について提案されてきた従来法の位置づけを整理するため，図 2･1, 2･2 に音声波形とその音声波形のパワースペクトルを示す．音声波形は，基本周期 T_0 で声帯振動が繰り返され，パワースペクトルは，f_0 Hz の基本波を示すピークに加え，その整数倍にもピークを持つ調波構造となる．したがって，基本周波数を推定する場合，音声波形に着目すると図 2･1 における T_0 を求める問題として扱われ，パワースペクトルに着目すると図 2･2 における基本波の周波数 f_0 を求める問題として扱われる．

(1) 時間波形に着目した方法

時間波形における性質を利用した方法では，信号の相関を用いる方法[4)]が一般的である．音声の基本周期が T_0 であれば，音声波形の自己相関，あるいは相互相関を求めると，T_0 の整数倍で高い相関値を示し，それ以外の時刻では低い相関値を示す．そのため，時刻 0 のピークを除いた最も早い時間のピーク時刻が T_0 となる．

相関に基づく方法を用いる場合，誤ったピークを選択することによる推定誤りが問題となる．一般的には，T_0 の整数倍以外に生じる不要なピークを除外するための閾値を設け，閾値を上回るピークの中から最も早い時間のピークを T_0 とする．しかしながら，声帯振動は，生じる時間間隔，および声帯振動波形がたとえ短時間であっても微細に変化しているため，目的とする時刻のピークが検出されない，あるいは T_0 以外，特に T_0 の整数倍の時刻のピークを誤検出する問題が起こりうる．

近年では，音声の相互相関関数を基準とし，不要なピークの低減や不要な演算を削減することにより，高速かつ高精度に基本周波数を推定可能な推定法 YIN [5)]が提案されている．YIN

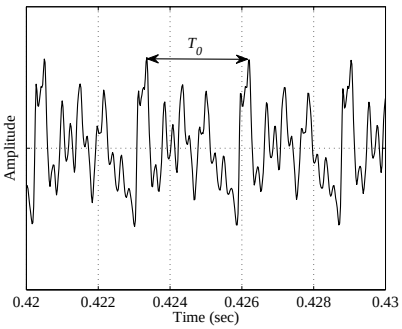

図 2·1　時間波形における基本周期．T_0 の逆数が f_0 となる．

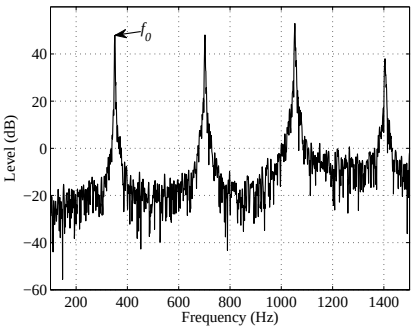

図 2·2　図 2·1 のパワースペクトル．最も低い周波数のピークが f_0 となる．

による基本周波数推定では，YIN の提案された 2002 年以前の従来法と比較して誤差を 1/3 以下に低減できることが文献 5) により示されている．

(2) パワースペクトルに着目した方法

周期信号のパワースペクトルは，f_0 の整数倍にピークを持つため，パワースペクトルの最も低いピークの周波数を抽出することで，基本周波数が推定できる．あるいは，f_0 の整数倍にピークを有する調波構造に着目し，調波構造のピーク間隔を推定することでも基本周波数を推定できる．パワースペクトルを基準とした推定法では，調音フィルタに起因する影響がパワースペクトルに混在するため，その影響を除去するための方法が要求される．

ケプストラム[6, 7]は，対数パワースペクトルを逆フーリエ変換することで得られ，ケフレンシーと呼ばれる時間を単位とするパラメタである．調音フィルタによる影響を分離可能であることから，ケプストラムは，基本周波数推定だけではなくスペクトル包絡推定にも利用される音声分析の代表的なパラメタといえる．音声のケプストラムを求めた場合，調音フィルタに起因するケプストラムは低次のケフレンシーに集中し，基本周波数に起因するピークが，調音フィルタに起因する成分よりも高次のケフレンシーである時刻 T_0 に生じる．そのため，高次ケフレンシーに存在するピークを抽出し，その逆数を求めることで基本周波数を推定可能である．

基本周波数が高い場合，基本周波数に起因するピークがケプストラムの低次に生じることとなる．相関を用いた方法と同様にケフレンシー軸における T_0 以外のピークを誤検出することに加え，調音フィルタのケプストラムが原因で正しいピークを抽出できないことが問題点となる．

2008 年に提案された SWIPE′ [8]は，パワースペクトルの調波構造に着目し，誤差を低減する様々な工夫を施すことで高い推定精度を達成する方法である．誤差を低減させる処理に複雑な演算が必要であることから計算コストは大きいが，YIN と比較した場合推定誤差をさらに低減することができる．

(3) その他の特徴量を用いた方法

これらの方法以外にも，より高精度の基本周波数推定を達成するため，様々な方法が提案されている．例えば，基本波をフィルタリングで取り出す方法[9)]，瞬時周波数を用いた方法[10)]や，ウェーブレット変換を用いた方法[11)]等が提案されている．この他にも，声帯振動時刻を直接検出する方法[12)]を用いることで，その間隔を直接求めることも可能である．2005 年に提案された NDF[13)]は，既存の複数の特徴量を抽出し，基本周波数の時間軌跡が滑らかになるよう前後の基本周波数を修正する後処理により推定精度を大きく改善している．

これらの方法は，高い精度を追求するため，計算コストに関する検討が行われていないのが現状である．近年では，音声や歌唱を加工することが可能なソフトウェアも実用化されているが，それらの音声分析においては，計算時間を可能な限り低減する必要がある．そこで本稿では，歌唱の分析や合成への応用を目的とした高速かつ高精度な基本周波数推定法[14)]を紹介する．

2-2-2 歌唱の分析・合成を目的とした基本周波数推定

歌唱分析・合成を目的としたコンテンツ制作を行う場合，そのコンテンツに用いられる音声は，防音室やレコーディングスタジオ等，背景雑音の少ない環境で録音される場合が多い．特に，歌唱合成等をコンテンツ制作現場で利用することを考慮すると，雑音がほぼ存在しない長時間の音声を対象として，正確な f_0 を可能な限り高速に推定することが望まれる．

文献 14) では，この目的を満たすため，分析対象とする音声を低域雑音を含まない音声に限定し，高速かつ高精度な基本周波数推定法が提案されている．この方法は，図 2･2 における最も低い周波数の基本波を低域通過フィルタにより抽出し，基本波の周波数を時間波形から計算する簡素な方法である．基本波検出に基づいて基本周波数を推定する場合，低域通過フィルタのカットオフ周波数は，推定時には未知である f_0 以上，その整数倍のピークの最小値となる $2f_0$ 以下に設定することが要求される．基本波をフィルタリングで取り出す従来法[9)]では，事前に別の方法で基本周波数候補を推定し，その近辺を抽出するフィルタリングが行われていた．文献 14) の方法では，f_0 の仮定をせず，以下に示される 3 つのステップにより基本周波数の推定を行う．

(1) ステップ 1: 低域通過フィルタによるフィルタリング

基本波の抽出は，最適な 1 つのカットオフ周波数を有する単一のフィルタではなく，低域から高域まで様々なカットオフ周波数を有する複数のフィルタ群により行われる．ステップ 1 では，様々なカットオフ周波数を有する低域通過フィルタ群により信号全体を処理する．フィルタ数に応じて計算コストは増大するが，このフィルタリングは波形全体に対する処理であり，フレーム単位で推定を行う従来法よりも高速な処理が可能である．音声処理には高い時間分解能が必要なため，低域通過フィルタには，カットオフ周波数以上の周波数のエネルギーを充分に抑圧できることだけではなく，フィルタ長が短く有限の時間で振幅が 0 に収束することが要求される．

(2) ステップ 2: 基本波らしさの計算

フィルタリングにより基本波のみが抽出された場合，その時間波形は周期が T_0 の正弦波となる．不要なピークを含む，あるいは基本波を含まない場合は，正弦波とは異なる波形となる．したがって，フィルタリングにより基本波が得られているか否かは，時間波形がどの

程度正弦波に近いのかを評価すれば良い.

この方法では，波形が正弦波の場合，信号のピークの間隔，谷の間隔，正から負のゼロ交差の間隔，負から正のゼロ交差の間隔が全て等しくなることに着目する．抽出された波形が正弦波に近いほど 4 つの間隔も等しくなるため，その標準偏差は 0 に近づくこととなる．基本波らしさは，4 つの間隔の平均を f_{ave}，標準偏差を f_{std} とした場合，$\exp(-f_{std}/f_{ave})$ により与えられる．基本波らしさは，0 から 1 の値を示し，1 に近いほど高精度に基本波を検出されたといえる．また，f_{ave} がその信号，その時刻における基本周波数の候補となる．

(3) ステップ 3: 基本波らしさに基づく最終的な基本周波数の選定

ステップ 2 により，フィルタリングされた各波形の基本周波数候補と基本波らしさが計算される．ステップ 3 では，全ての候補から，各時刻における最終的な基本周波数を選定する．ただし，低域通過フィルタの条件より，以下の条件のいずれかを満たす候補は除外される．

- フィルタリングに用いられたカットオフ周波数の下限を下回る候補，上限を上回る候補
- 低域通過フィルタの通過域以外の周波数に存在する候補
- 低域通過フィルタのカットオフ周波数の半分以下の周波数に存在する候補

この選定処理で残った候補から，最も基本波らしさの大きい候補を，最終的な候補とする．

本方法は，低域に雑音が存在する音声に対する推定は困難であるが，低域の雑音が存在しない音声の場合，SWIPE′ や NDF と実質的に同等の性能を達成しつつ，計算時間を SWIPE′ の 1/42, NDF の 1/80 にまで低減可能である．

2-2-3 今後の展望

基本周波数は，音声の主要な要素であり，その高精度な抽出は，学術的だけではなく，歌唱合成ツールなど産業的な価値も高い研究テーマである．特に，商用化を目指す場合，計算コストに関する問題があるため，限られた計算時間でより高精度な推定性能を達成するための工夫[15]が求められる．また，CGM が一般化した現在，一般ユーザ向けの技術提供も重要な課題である*.

■参考文献

1) 中野倫靖，後藤真孝，平賀譲，“楽譜情報を用いない歌唱力自動評価手法，” 情報処理学会論文誌， Vol. 48, No. 1, pp. 227-236, 2007.
2) 齋藤毅，後藤真孝，鵜木祐史，赤木正人，“SingBySpeaking：歌声知覚に重要な音響特徴を制御して話声を歌声に変換するシステム，” 情報処理学会研究報告，2008-MUS-74, pp. 25-32, 2008.
3) W. Hess, “Pitch determination of speech signals,” Springer-Verlag, Berlin, 1983.
4) M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, H.J. Manley, “Average magnitude difference function pitch extractor,” IEEE Transactions on acoustic, speech, and signal processing, vol.ASSP-22, no.5, 1974.
5) 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.
6) 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.

* 筆者により提案された音声分析変換合成システム WORLD[15]は，UTAU[16]用に実装され，様々な楽曲制作者により利用されている．

7) A.M. Noll, “Cepstrum pitch determination,” J. Acoust. Soc. Am., vol.41, no.2, pp.293-309, 1967.
8) 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.
9) 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.
10) 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.
11) 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.
12) K.S.R. Murty and B. Yegnanarayana, “Epoch extraction from speech signals,” IEEE Transactions on audio, speech and language processing, vol.16, no.8, 2008.
13) H. Kawahara, A. Cheveigné, H. Banno, T. Takahashi and T. Irino, “Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT,” Proc. Interspeech2005, pp.537-540, 2005.
14) 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.
15) 森勢将雅，中野皓太，西浦敬信, “歌唱合成システムの実現を目的とした高品質音声分析合成法の提案，” 電子情報通信学会 応用音響研究会, vol.110, no.71, pp.89-94, 2010.
16) “歌声合成ソフトウェア UTAU, ” http://utau2008.web.fc2.com/index.html