2008－MUS－76 (17)
2008／8／7

零周波数フィルタ信号に基づく基本周波数抽出法のTANDEM-STRAIGHTへの応用について

河原 英紀 †, 森勢 将雅 ††, 坂野 秀樹 ‡, 板垣 英恵 † 大西 壮登 † 西村 竜一 † 入野 俊夫 †

† 和歌山大学システム工学部/システム工学研究科 †† 関西学院大学理工学部 ‡ 名城大学理工学部

Yegnanarayana らは、インド語の CV 連鎖における破裂子音の分析を目的として、零周波数に 4 重の極を有するフィルタと局所的平均値を除去する FIR フィルタを組み合わせ、声帯の動作に関連するイベントを抽出する方法を提案した [1]。ここでは、TANDEM-STRAIGHT および real time STRAIGHT への応用を狙い、追試および幾つかの評価を行った。その結果、この方法は、laptop PC 上の Matlab を用いた実装でも実時間の 1/7 で基本周波数を抽出することができること、最新の方法にはやや劣るものの十分に実用になる gross error である 0.55%が達成されること、瞬時周波数に基づく方法と同等の結果を、1/3 程度の持続時間という高い時間分解能で求められることが示された。

F0 extraction based on the zero frequency filtered signal method and its application to TANDEM-STRAIGHT

Hideki KAWAHARA † Masanori MORISE †† Hideki BANNO ‡ Hanae ITAGAKI ‡‡ Masato ONISHI ‡‡
Ryuichi NISIMURA † Toshio IRINO †

† Faculty of Systems Engineering, Wakayama University

†† School of Science and Technology, Kwansei Gakuin University

‡ Faculty of Science and Technology, Meijyo University

‡‡ Graduate School of Science and Technology, Wakayama University

An event based f0 extraction method based on so called zero frequency filtering method was proposed by Yegnanarayana for representing Indian stop consonants [1]. The proposed method uses unstable IIR filters that place four poles at zero frequency and at the same time employs local mean subtracting filters to stabilize its output. This simple method was reported to run extremely fast and has comparative performance with existing F0 extractors. This article reports on a follow-up implementation of the method and evaluations and investigations for its performance and characteristics having its applicability to TANDEM-STRAIGHT and real time STRAIGHT in mind. The results indicated that the proposed method runs 7 times faster than real time with Matlab implementation on a standard laptop PC. It was also found that the gross error rate was 0.55% which is somewhat worse than the most recent methods but still reasonably high for practical applications. Finally, temporal resolution finer (namely 1/3) than instantaneous frequency based methods was also demonstrated.

1 はじめに

STRAIGHT [2, 3] に基づくモーフィング [4] 等、歌唱音声の柔軟な操作のためには、基本周波数をはじめとする音源情報の正確な抽出が必要となる。朗読音声や普通の歌声での母音定常部等、周期性が明瞭に認められる場合には、最近提案されているアルゴリズム [6, 5, 7] が、実用上問題の無い性能を示す。しかし、強い表情の感情音声や歌唱、母音の開始や終了部など、声帯振動が不規則になる部分であ、様々な問題が生ずる。

初期推定の結果に対する後処理 [8] や、局所的な繰り返し構造を抽出する方法 [9] は、これらへの対策の例である。しかし、これらは一般に多くの計算量を必要とする。非常に高い精度の領域では、『基本周波数』という概念そのものの妥当性が怪しくなる。また、ライブやインタラクティブシステムへの応用では、リアルタイム処理に適した高速で遅延の少ないアルゴリズムが必要とされる。

Yegnanarayana らは、音声の駆動の重要な原因である声帯の動作そのものの抽出を目的として、零

周波数フィルタリングに基づく方法[1]を提案した。以下では、その追試と、TANDEM-STRAIGHT[10]への応用について説明する。

2 声帯音源の分析

破裂の時点から声帯が振動を開始するまでのVOT（Voice Onset Time）や、付随して生ずる気息性の雑音のタイミングと変動のパターンは、破裂子音を区別するための知覚的な手掛かりである。零周波数フィルタリングに基づく方法は、これらの分析に必要な、声帯の動作の正確な抽出を行うために提案された[1]。

2.1 零周波数フィルタリング

文献[1]では、以下の手順によりフィルタ出力を求めている。

STEP 1 入力音声波形 $s[n]$ から差分信号 $x[n]$ を作成する。

$$x[n] = s[n] - s[n-1] \tag{1}$$

STEP 2 零周波数に二重の極のあるIIRフィルタを用いて以下のように処理する。

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \tag{2}$$

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n] \tag{3}$$

ここで $a_1 = -2$, $a_2 = 1$ である。これは、4重の積分と等価である。

STEP 4 10msの移動平均フィルタを用いて局所的な平均値を求め、除去する。（この処理は、零周波数に二重の零を配置することに相当する。）

2.1.1 処理の変更

この記述のままではトレンド成分が除去されず、文献の動作を再現することができなかった。ここでは、STEP 1を省き、STEP 4を繰り返すこととした。以下では、この変更を加えたものを用いる。

2.2 分析例

図 1 に、男性の発声した日本語の母音連鎖/aiueo/の分析例を示す。下の図は、母音開始部分での音声波形（赤実線）とフィルタ出力（青実線）を示す。音声波形が周期的に急に大きな振動を開始する位置は、声門が閉止する時点に対応している。フィルタ出力は、ほぼ、この時点において負

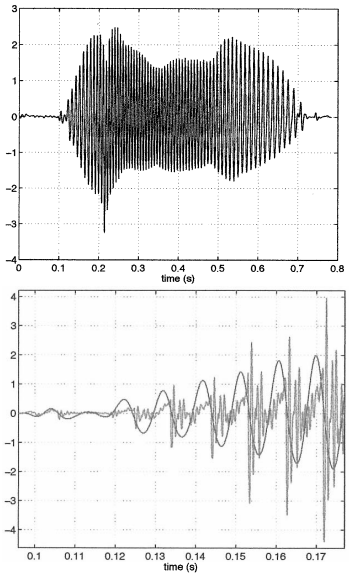

Fig. 1 Filtered waveform (upper). Filtered waveform and the original signal (lower). The red line in the lower plot represents the original signal of a Japanese vowel sequence /aiueo/ spoken by a male speaker.

の方向に零を通過する。従って、隣接する零交差間隔を測ることにより、声門閉止が繰り返される間隔（基本周期に対応）が求められ、その逆数として基本周波数が求められる。

なお、この試料は、無指向性の圧力型コンデンサマイクロフォンを用い、低域での位相回転の少ない計測用のプリアンプを介して録音されている。単一指向性のマイクで近接効果補償用の低域カットを行った場合には、マイク本体とフィルタの双方で基本波の位相が大きく回転してしまうため、このような対応関係を保証することはできない。

声門の運動のタイミングを議論する場合には、このように収録系での位相に配慮することが必要となる。しかし、基本周波数の抽出を目的とするのであれば、同一の位相に戻る条件の再現性が保証できるだけで良い。以下では、基本周波数のみに注目することとする。

3 基本周波数抽出への応用

局所的な平均を求めるための区間長がこの方法での唯一の調整可能なパラメタとなる。文献では、この区間長を分析対象の音声の平均基本周波数の逆数に設定するとよいとの記述がある。ここでは、

以下のように処理系を構成して最適な区間長を決定し、基本周波数抽出を行うこととした。

3.1 処理系の構成

分析以前には平均基本周波数は未知である。そこで、まず 40Hz から 806Hz までの 1/3 オクターブ毎に基本周波数を仮定して複数の分析を並行して行う。次いで、その結果を比較して平均値除去のための適切な区間長を決定し、最終的な分析を行うこととした。

分析結果の比較には、求められた平均基本周波数を用いた。信号中の基本周波数成分（基本波成分）が十分に強ければ、求められた基本周波数の平均値が、設定した基本周波数に依存せずに一定の値を示す範囲が広くなることが期待できる。また、零交差間隔も、基本波成分が支配的であれば、隣接する周期での間隔の変動量が少なくなることが期待できる。これらを合成した評価値を最小とする分析結果を採用し、その平均基本周波数を基準として、平均値除去のための適切な区間長を決定する。この区間長を平均基本周期の何倍とするかを指標として、後述の評価を行った。

3.1.1 各部の動作例

以下では、後述のデータベース [11] に収録されている女性の発声した「コーヒーにミルクを入れますか？」という文章音声を用い、各部の動作を説明する。データベースには、EGG（Electroglottograph）の視察により求められた有声区間も併せて記録されている。

図 2 に、区間長の設定値により、求められた基本周波数の平均値がどのように変化するかを示す。ここでは、区間長を設定した基本周波数の逆数とすることとし、横軸を設定した基本周波数とした。設定した基本周波数が 200Hz から 600Hz の範囲において、求められた平均基本周波数は、ほぼ250Hz となっている。中段のヒストグラムでは、この様子が分かり易い。下段の図は、隣接する周期での間隔の変動量を表す。ここでも、同様の設定範囲で変動量が少なくなっていることが分かる。

このようにして、決めた基本周波数の設定値を用いて求めた基本周波数の例を図 3 に示す。基本周波数は、零周波数フィルタの出力における隣接する零交差の間隔として求められる。図中の赤線は、視察により求めた有声音区間であり、データベースの第三チャンネルに記録されている。下の図は、文献の方法で求められた零周波数フィルタ信号の零交差時刻における信号の傾斜を示す。視察による結果との比較により、傾斜情報が、有声

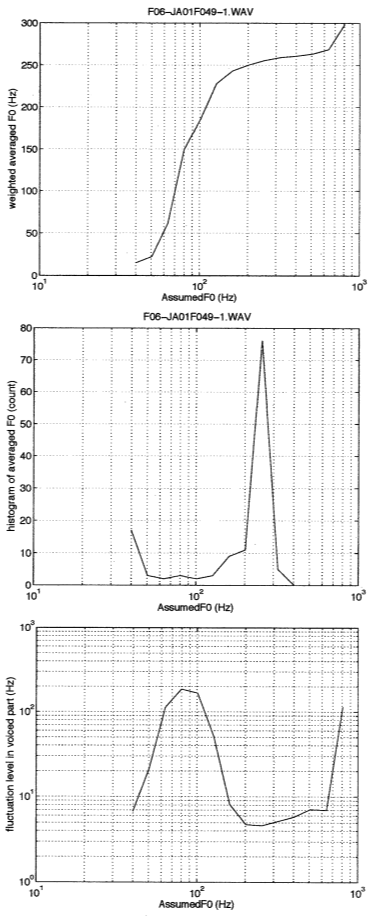

Fig. 2 Weighted average F0 for assumed F0 (top), histogram of estimated F0 values (middle) and averaged cycle-by-cycle F0 fluctuations for assumed F0 (bottom)

／無声判定の有用な手掛りなることが分かる。

4 評価

音声と EGG 信号を同時収録したデータベース [11] を用い，今回実装した方法の性能を調べた。データベースには、男女各 14 名の発声した 30 文章からなる合計 840 文章が収録されている。予備検討により、今回の方法は、EGG 信号の分析では（微分波形、二階微分波形を用いても）多くの誤りを生ずることが明らかとなった。そこで、後処理を多用した F0 抽出法 [8] により EGG 信号を分析したものを正解であるとして以下の評価を進めた。

今回実装した方法のパラメタは、局所的な平均

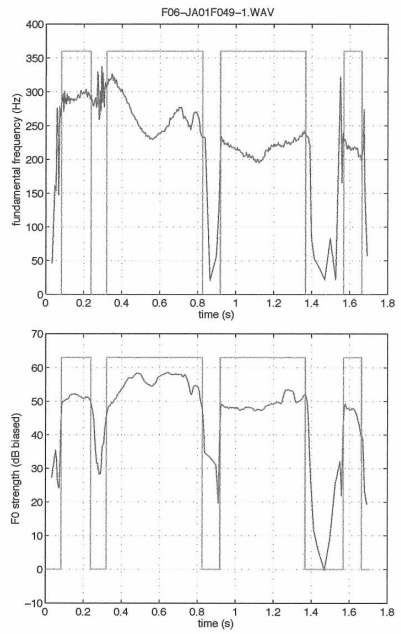

Fig. 3 Estimated F0. Red lines represent manually identified voiced region (upper). Zero-crossing speed of the filtered signal (lower)

を除去するための移動平均フィルタの窓長である。窓長の逆数が、最初の零点の周波数に対応する。性能は、この零の周波数を平均基本周波数の何倍とするかに依存する。ここでは、平均基本周波数の0.5倍から1.5倍までの範囲で調べた。

図4に、正解で正規化した推定値のヒストグラムを示す。上は全体、下は100%の近傍の様子を示す。設定値の係数を大きくすると100%付近の分布は鋭くなる。しかし、この傾向は、係数が1以上では飽和する。しかも、上の図から分かるように、高い周波数側に生ずる誤差が増加する。

図5の上の図に、設定値と gross error の関係を示す。図より、男女の平均で見ると、設定値を基本周波数の平均値と一致させた場合に最も gross error が少なくなる。この結果は、文献[1]の主張と整合する。下の図より、840文章中、240個以上の文において gross error が零であることが分かる。最悪の場合の gross error は、11%である。

最悪の gross error を示す音声と、その分析結果を図6に示す。最下段に示した TANDEM-STRAIGHT による分析結果と比較することにより、収録に用いたマイクの急峻な低域遮断特性に

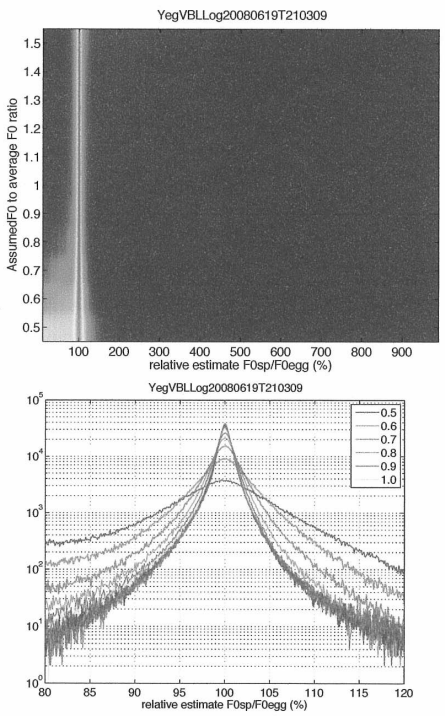

Fig. 4 Image plot of the histogram of the relative F0 values in terms of EGG-based F0 as the reference. The vertical axis represents ratio of the averaging duration for localized mean value removal to the averaged fundamental period (upper). Magnified view of the histogram around 100% (lower)

Table 1 Gross error rate comparison with other F0 extractors. (Excerpts from the reference on NDFF[8]. Yegna: propsed method, NDF: current optional STRAIGHT F0 extractor with heavy post processing, TEMPO: default STRAIGHT F0 extractor. Refer also to the article on YIN[5] for details)

Method	DB1	DB2
Yegna	0.55	–
NDF (STRAIGHT)	0.09	0.35
TAMPO (STRAIGHT)	0.77	2.8
YIN	0.29	1.4
ac (autocorrelation)	2.7	7.3
fxcep (cepstrum)	4.5	12.5

よる基本波レベルの低下が、この大きなエラーの原因であることが分かる。

表1に、相対的なF0の推定誤差が20%以上となる割合として定義される gross error を、幾つかの

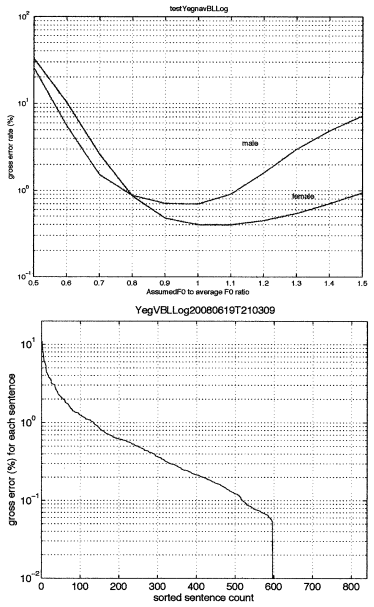

Fig. 5 Averaged gross error based on each sentence gross error. Blue line represents male and the red line represents female results (upper). Ordered gross error for each sentence (lower)

F0 抽出法と比較して示す。提案した方法以外は、文献 [8] の結果を再掲した。最も優れた結果を与えるものではないが、実用上問題の無いレベルにあると言えよう。

なお、提案した方法は、非常に高速である。MacBookPro (2.2 GHz Intel Core 2 Duo) 上の Matlab を用いて実装したプログラムは、図 3 に示した 1.7 秒の音声からの F0 抽出を 0.23 秒で実行した。実時間の約 1/7 である。この実装では、区間長の選択と併せて、15 回の F0 抽出が行われている。すなわち、区間長が既知である場合には、Matlab を用いた実装であっても、実時間の 1/100 で基本周波数を求めることができる。real time STRAIGHT 等、速度が必要とされる応用に適した性質である。

なお、基本波が選択されているのであれば、零交差の代わりに瞬時周波数を求めることで基本周波数を抽出できそうである。しかし、瞬時周波数の安定な抽出のためには、十分な長さの窓関数により基本波のみが選択されるようにする必要がある。図 7 に、本方法の処理に対応するインパルス応答と、直交位相信号を用いて基本波のみを選択する場合に必要な包絡の例を示し、図 8 に、それ

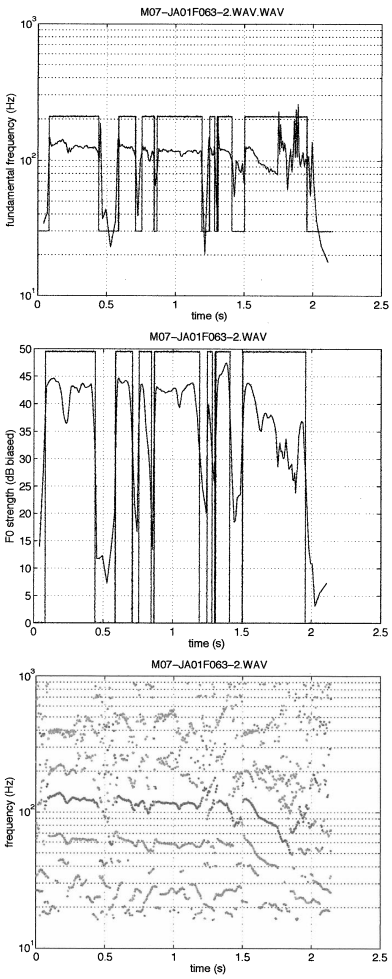

Fig. 6 Estimated F0 of the sentence with the worst gross error 11% (upper). Zero-crossing speed of the filtered signal (middle). Extracted F0 candidates by TANDEM-STRAIGHT. Blue dots represent the best candidates (lower).

らを用いた基本周波数の抽出結果を、本方法と比較して示す。

図 8 より、本方法と同程度の結果を得るためには、基本周期の 3〜5 倍の長さの窓関数を用い、しかも適切な中心周波数を設定しなければならないことが分かる。図 7 に示した関数に対応する持続時間は、本法では、2.28 ms、$3T_0$ の窓関数の場合は、4.78 ms、$5T_0$ の場合は、7.97 ms である。本方法が高い時間分解能を有していることが分かる。この

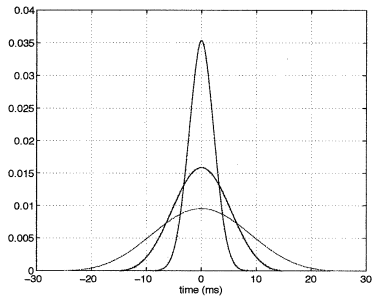

Fig. 7 Equivalent filter impulse response of the proposed method (blue) and envelopes of quadrature signals for instantaneous frequency calculation. Blackman window is used to define the envelopes. (red: length with $3T_0$ and green: length with $3T_0$, where $T_0 = 10$ ms is assumed.)

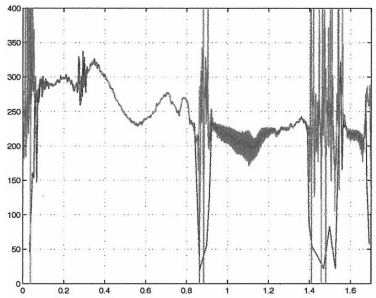

Fig. 8 Instantaneous frequencies of filter outputs and F0 by the proposed method. (blue) proposed method, (red) quadrature signal with Blackman window envelope (length: $3T_0$), and (green) quadrature signal with Blackman window envelope (length: $5T_0$)

高い時間分解能は、TANDEM-STRAIGHT の音源情報抽出部分を拡張する上で非常に有用である。

5 まとめ

Yegnanarayana らにより提案された零周波数フィルタリングに基づく基本周波数抽出法を実装し、性能評価と TANDEM-STRAIGHT への応用可能性の検討を行った。その結果、本方法が、実用上充分な性能を有すること、高速性および高い時間分解能という有用な特徴を有することが示された。

謝辞

貴重な時間を割いて議論して頂いた産総研の後藤真孝 主任研究員、応用を試みて頂いた京都大学の高橋徹 GCOE 助教に感謝します。なお、本研究の一部は、科学技術振興機構による戦略的創造研究推進事業のデジタルメディア領域 CrestMuse プロジェクトと、科学技術研究費補助金 (A)19200017 による支援を得て行われた。

参考文献

1) B. Yegnanarayana , K. Sri Rama Murty and S. Rajendran: Analysis of Stop Consonants in Indian Languages Using Excitation Source Information in Speech Signal, Proc. ITRW Aalborg, Denmark, 2008.
2) H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction", *Speech Communication*, 27(3–4), pp. 187–207 (1999).
3) 河原, "Vocoder のもう一つの可能性を探る：音声分析変換合成システム STRAIGHT の背景と展開," 日本音響学会誌, Vol.63, No.8, pp.442–449 (2007).
4) H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation", *ICASSP'2003*, I, pp. 256–259 (2003).
5) A. de Cheveigné, H. Kawahra, "YIN, a fundamental frequency estimator for speech and music", *Journal of the Acoustical Society of America*, Vol.111, No.4, pp.1917–1930 (2002).
6) Kawahara, H., Katayose, H., de Cheveigné, A., and Patterson, R. D., "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," Proc. EUROSPEECH'99, 6: 2781–2784, 1999.
7) A. Camacho: SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech And Music, Ph.D. Thesis, University of Florida, 116p., 2007.
8) Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., and Irino, T., "Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT," Proc. Interspeech'2005, 537–540, 2005.
9) Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Hideki Banno, Toshio Irino: A unified approach for F0 extraction and aperiodicity estimation based on a temporally stable power spectral representation, Proc. ITRW Aalborg, Denmark, 2008.
10) Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., "TANDEM–STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," ICASSP 2008, Las Vegas, pp.3933-3936, 2008.
11) 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, 電子情報通信学会誌, D-II, J83-D-II, 11, pp.2077–2086, 2000.