Vol.2012-MUS-95 No.4
2012/6/2

音源およびスペクトル包絡の時間的微細構造の加工と歌唱音声の印象への影響について

河原 英紀[1,a)]　森勢 将雅[2,b)]　西村 竜一[1,c)]　入野 俊夫[1,d)]

概要：シャウトやデスボイスなどの激しい表現は、ポピュラー歌唱で広く用いられている。これらを適切に分析、再現、制御する方法を明らかにすることは、歌唱合成システムに豊かな表現力を与えるために解決すべき重要な課題である。本報告では、まず、新たに開発した高い時間分解能を有する基本周波数抽出法とそれに基づく TANDEM-STRAIGHT により、様々な歌唱音声を分析した結果について報告する。分析結果は、激しい表現にいおいて、70 Hz 付近に 20 dB 程度の高さのピークを有する高速の（基本周波数の）周波数変調と、同様に、高速の（スペクトル包絡の）振幅変調が存在することを示した。このような高速の変調の存在は、これまでにはっきりとは報告されていない。予備的な実験により、それらの高速の変調を加工することにより、発声の声区と努力の印象を保ったまま、シャウトなどの歌唱表現の強さ（生々しさ）を制御できる可能性が示された。

キーワード：歌唱音声、周波数変調、振幅変調、スペクトル包絡、基本周波数

Manipulation of temporal fine structures on excitation source and spectral envelope of singing voices and their effects on perceived impression

Kawahara Hideki[1,a)]　Morise Masanori[2,b)]　Nisimura Ryuichi[1,c)]　Irino Toshio[1,d)]

Abstract: Strong expressions such as “shout” and “death voice” are common in popular singing. However, current singing synthesis systems are not good at handling these strong expressions and are not capable of using them to expand their limit of expressiveness. This is the topic this article tries to address. A set of singing voice analysis tests was conducted using our newly developed F0 extraction method, which has high temporal resolution and is light-weighted, and TANDEM-STRAIGHT for spectral envelope analyses. This test revealed that expressive singing voices consist of high-speed frequency as well as amplitude modulations in F0 and spectral envelope respectively. In one typical case, about 20 dB higher modulation frequency spectral peak was found around 70 Hz for expressive performance than that of normal performance. Preliminary tests suggested that selective control of “expressiveness” can be implemented by manipulating these high-speed modulations while preserving vocal register and effort intact.

Keywords: singing voice, frequency modulation, amplitude modulation, spectral envelope, fundamental frequency

[1] 和歌山大学
Wakayama University, Wakayama, 640–8510, Japan
[2] 立命館大学
Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan
[a)] kawahara@sys.wakayama-u.ac.jp
[b)] morise@fc.ritsumei.ac.jp
[c)] nisimura@sys.wakayama-u.ac.jp
[d)] irino@sys.wakayama-u.ac.jp

1. はじめに

印象的な歌唱の音声には、様々な強い表現が含まれる。ジャンルによっては、美しい響きの歌声の表現が意図的に避けられ、シャウトなどの強い表現が全般に亘って用いられている場合すらある。そのような歌唱音声の分析に、基

本周波数軌跡の滑らかさやスペクトル包絡の定常性を仮定する分析法を用いることは、適切ではない。ここでは、滑らかさや定常性の仮定に依存しない分析法を用いることで、強い表現に関わる音声の物理的特徴を明らかにし、それらの特徴の加工により強い表現に関わる印象を操作する可能性について、検討した結果を報告する。

2. 基本周期に適応した分析

有声音における周期的駆動を、背景となる時間周波数表現を標本化する手段として解釈することにより、STRAIGHT [1] および TANDEM-STRAIGHT [2] が導かれている。これらで用いられる分析のための時間窓長は、基本周期に比例して適応的に設定される。そのため、これらの方法で求められる基本周波数およびスペクトル包絡には、(窓とその後の処理による平滑化に起因する減衰はあるものの) 基本周波数を標本化周波数と見なした場合のナイキスト周波数までの変動が含まれていると考えて良い。TANDEM-STRAIGHT の時間分解能についての議論は、文献 [3] に譲り、ここでは高速で軽量な新たな基本周波数分析法について説明する。

2.1 基本波の対称性に基づく分析

Yegnanarayana らによる初期の方法 [4] から零周波数フィルタリング [5] に基づく方法まで、様々な音声の駆動情報の抽出法を比較した結果が報告されている [6]。その中で推奨されている零周波数フィルタリングに基づく方法は、(発明者らの主張とは異なるが) 低域フィルタを利用した基本波の選択と波形の繰返し間隔を測定する方法を組み合わせたところに本質があると理解することができる。

基本波の選択に基づく基本周波数抽出法には様々な提案 [7] があり、国内でもエネルギーオペレータを併用した方法 [8] が提案されている。しかし、これらの方法では基本波成分の選択に帯域通過フィルタが用いられており、時間分解能を大きく損なう原因となっている。また、抽出誤りを回避するために基本周波数の変化の滑らかさを仮定しているものが多い。この時間分解能の劣化の問題を、低域通過フィルタを用いることにより避け、高速に基本周波数を抽出する方法 [9] を提案して来た。また、波形の対称性に基づく指標を導入することで、直流バイアスに対する耐性と時間分解能を改良する方法 [10] を明らかにした。ここでは、波形の対称性に基づく方法の最近の改良を併せて説明し、歌唱音声への適用結果について報告する。

2.1.1 処理の概要

基本周波数の分析に際し、事前情報として基本周波数の値が未知であることを前提とする。基本波成分を選択するフィルタの設計に用いることのできる情報が無い場合には、基本周波数が存在する可能性のある周波数帯域を適切な密度で覆うことができるように、複数のフィルタを用意し、その中から適切なフィルタを選択することが必要となる。この選択の指標に、フィルタ後の波形の対称性を利用する。

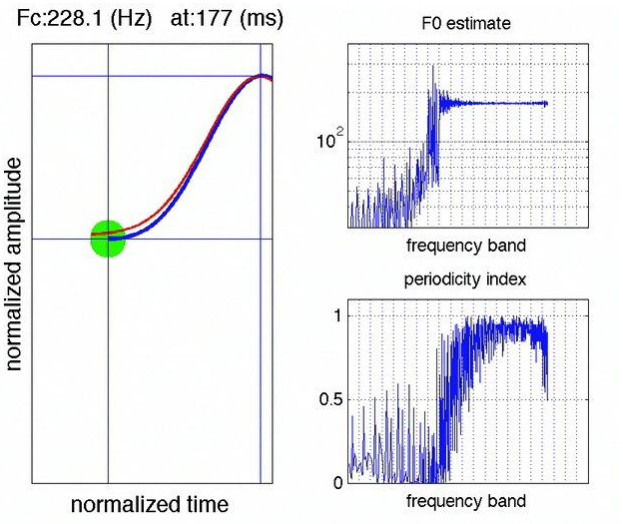


図 1 波形の対称性に基づく指標と、基本周波数抽出のためのフィルタ選択。左図に、極大点で時間方向に折り返した波形 (赤) と元の波形 (青) のズレを緑のマーカーの大きさで示す。右下には、このズレを評価し、0 と 1 の間の値をとる指標に変換した値を横軸で表した遮断周波数の関数として示す。右上は、低域通過フィルタの遮断周波数とそれぞれのフィルタの出力波形から求められた基本周波数を示す。

Fig. 1 Symmetry-based filter selection. Left plot illustrates deviation from symmetry using green marker based on the original half cycle (blue) and the mirror image of the succeeding half cycle (red). This discrepancy is converted to relevance index (ranging from 0 to 1) and shown in the right bottom plot as a function of LPF cut-off frequency. The right top plot shows fundamental frequencies derived from the filter outputs.

基本波だけが選択されている場合には、波形の対称性が高いため、波形の極値を中心として波形を折り返すと、隣接する極値が重なる。逆に言えば、この折り返された極値間の距離の大きさが、フィルタ選択の不適切さを表すことになる。この距離には振幅変調に起因するものと、周波数変調に起因するものが含まれている。それらの適切な重み付けのために Minkowski 距離を用い、成分の相対的な値と、距離の指数を調整する。こうして選択されたフィルタの極値間の (折り返さない場合の) 時間間隔として基本周期 (その逆数として基本周波数) が求められる。

図 1 にフィルタ選択のための指標を説明する表示の一例を示す。ここでは、SNR 30 dB のパルス列が試験信号として用いられている。また、図 2 に、実際の音声 (男性話者による日本語の母音連鎖/aiueo/) の分析の際の動作例を示す。選択されたフィルタの遮断周波数の軌跡にある比較的大きな変動は、フィルタ出力の周波数が広い範囲でほぼ同一の値となっているため、求められる基本周波数にはほと

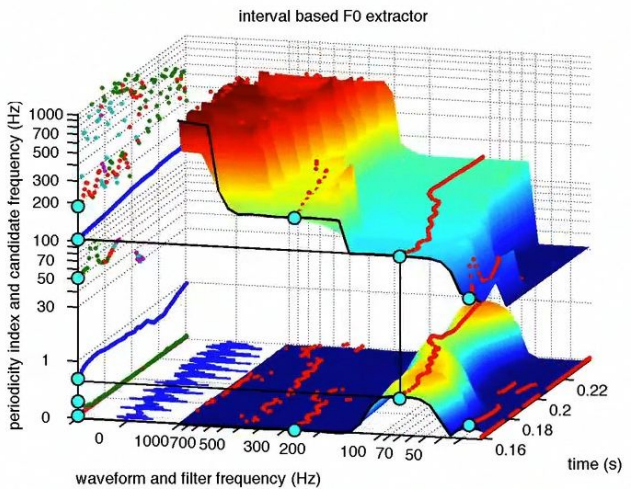


図 2 実音声での指標と基本周波数の抽出。音声は男性の発声した日本語の母音連鎖/aiueo/。左右の軸が低域通過フィルタの遮断周波数、前後の軸が時間を示す。下の彩色された曲面の高さは、0 と 1 の間の値をとる指標を示し、極大値が赤点により表示されている。最大の極大値を与える点の遮断周波数に基づいて彩色された上の曲面の値が読み出されて、基本周波数が求められる。左側の垂直面には、それぞれの極大値に対応する値が表示されている。下の平面の左端には、対応する音声波形が示されている。

Fig. 2 Symmetry-based filter selection for natural speech. The material is a Japanese vowel sequence /aiueo/ spoken by a male speaker. The horizontal axis represents cut-off frequency of LPF and the front-back axis represents time. The color mapped curved surface underneath represents the index of relevance. Red dots shows maxima locations of each frame and the most relevant location (cut-off frequency) is used to read the frequency value of the filter output from the upper colored surface. The left wall displays the values of maxima points and corresponding frequency values. The left most patch on the floor shows the corresponding waveform.

んど影響を与えないことが分かる。

この基本波成分の測定により求められた基本周波数は、単一の成分の情報のみに基づいているため、雑音による影響を大きく受ける。そのため、基本波成分の測定により求められた基本周波数を初期値とし、調波成分の瞬時周波数を利用した改良を繰返すことにより、精度の高い基本周波数を求める。なお、ここで用いる瞬時周波数は、TANDEM と同様の手法により、周期性に起因する変動を取り除いたもの [11] である。この修正により、自乗平均値で評価した推定誤差は、ほぼ 1/10 となる [10]。なお、初期値と修正値の差は、図 2 のように表示した場合には、重なってしまい区別できない程度である。基本周波数が 200 Hz の場合、変調周波数伝達特性の利得が-3 dB となる周波数を用いて時間分解能を表すと、初期推定値では 70 Hz、修正値では 50 Hz となる。これらは YIN [12] や SWIPE′ [13] など、広く使われている方法を大きく凌いでいる。

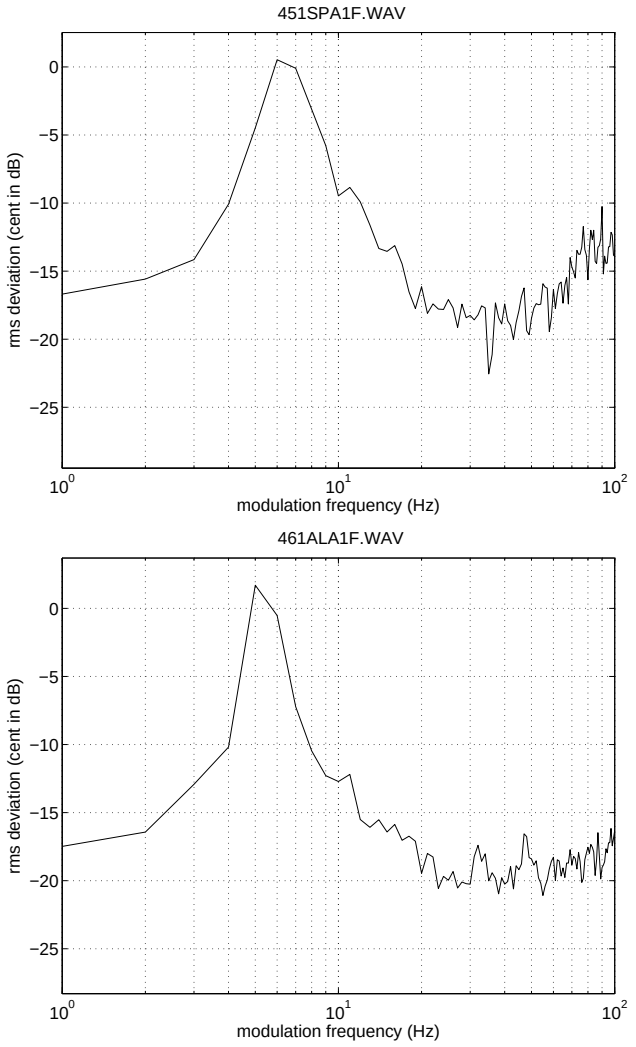


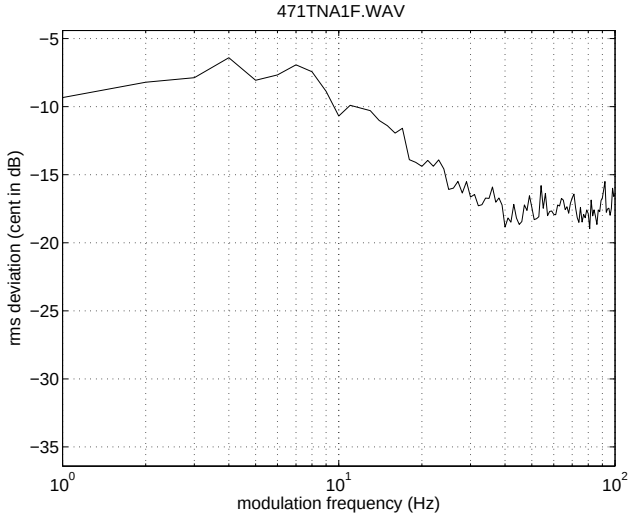
図 3 cent で表された基本周波数軌跡の差分信号による周波数変調スペクトル。上はソプラノ、下はアルトの歌唱の分析結果を示す。

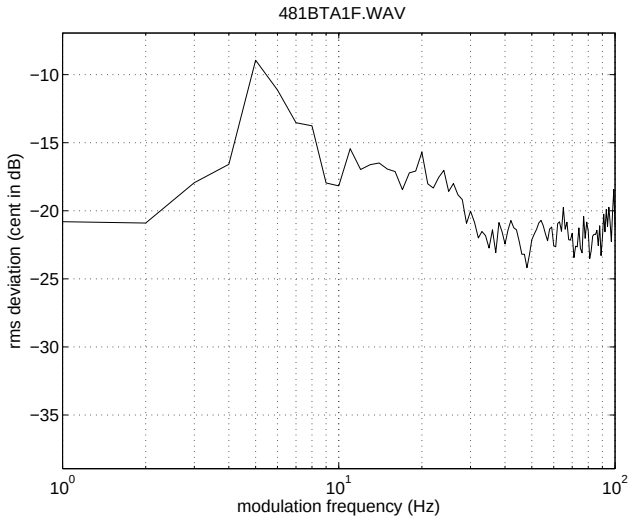
Fig. 3 Modulation power spectrum of differentiated fundamental frequency (represented in cent) for female singers. Top: soprano and Bottom:Alto.

3. 持続母音歌唱の分析

ここでは、まず、RWC 研究用音楽データベースの歌唱音声 [14] を対象として、基本周波数軌跡に含まれる変動を調べた結果について報告する。以下の分析ではフレーム周期 1 ms を用いている。基本周波数の軌跡は cent に変換した後、基本周波数が安定している区間を自動的に切出し、開始及び終了部分をそれぞれ 100 ms 取り除いた後、差分処理したものを分析の対象とした。また、区間内に抽出誤りが含まれている区間を分析対象から外した。

歌声のデータベースからビブラートを含むフォルテでの母音/a/の歌唱音声を選択し、分析した結果の抜粋を図 3〜図 5 に示す。クラシック歌手の歌唱に認められる 5 Hz から 7 Hz 付近の鋭いピークは、ビブラートによる。図 4 のピークが広がっているが、これは、この歌手の発声が不安





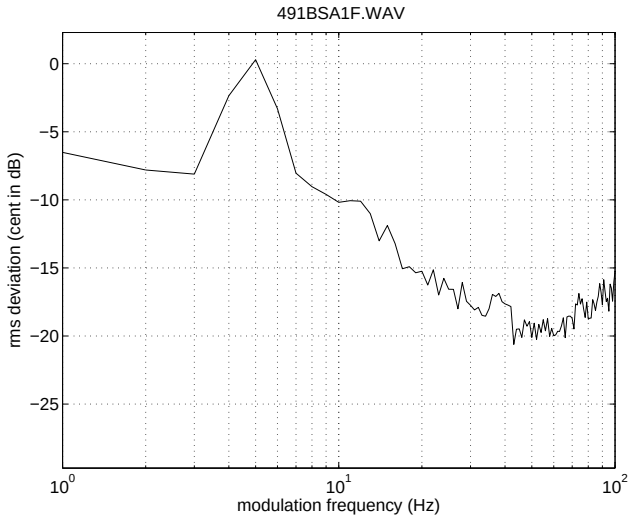
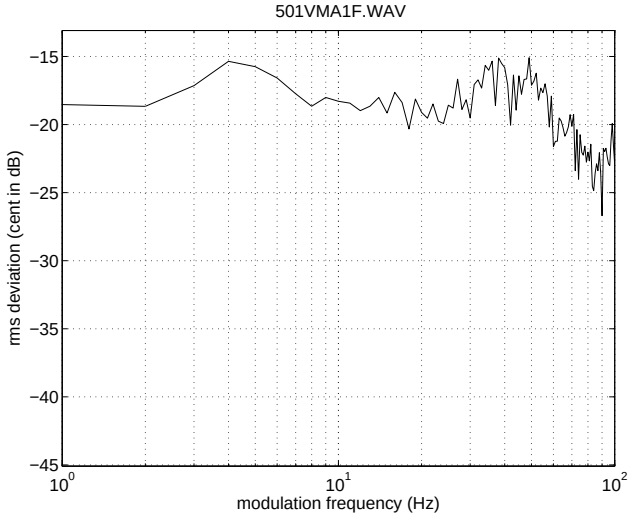
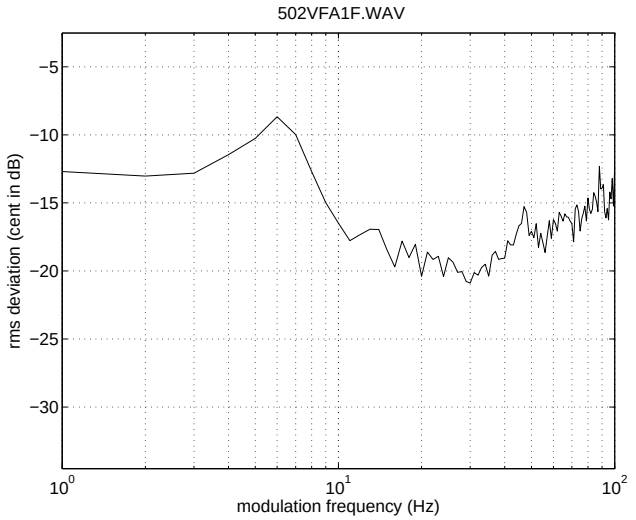


図 4 cent で表された基本周波数軌跡の差分信号による周波数変調スペクトル。上はテノール、中段はバリトン、下はバスの歌唱の分析結果を示す。

Fig. 4 Modulation power spectrum of differentiated fundamental frequency (represented in cent) for male singers. Top: tenor, Middle:baritone and Bottom:bass.





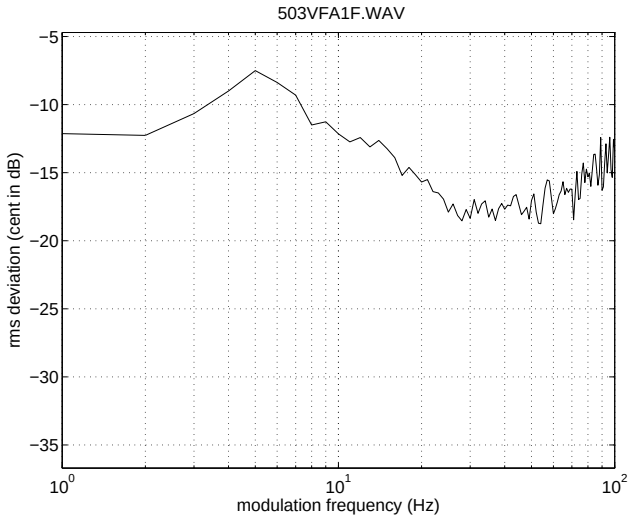


図 5 cent で表された基本周波数軌跡の差分信号による周波数変調スペクトル。R&B 系のポピュラー歌手の歌唱を収録。上は男性歌手、中段と下段は女性歌手の歌唱の分析結果を示す。

Fig. 5 Modulation power spectrum of differentiated fundamental frequency (represented in cent) for popular song singers (R&B). Top: male, Middle and Bottom:female.

定であることを反映したものと考えられる。

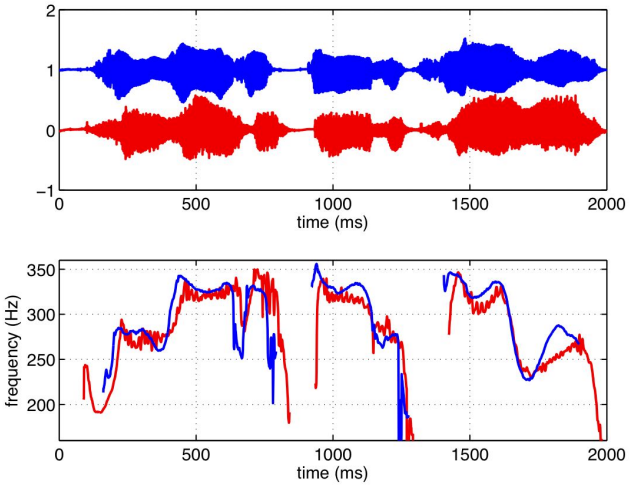


図 6 表現による歌唱音声の波形と基本周波数の変化。上段に波形、下段に基本周波数軌跡を示す。赤は表情豊かな演奏、青は無表情な演奏を示す。

Fig. 6 Waveforms and F0 trajectories of two singing expressions. Upper plot shows waveform and lower plot shows F0 trajectory. Red lines: expressive performance and Blue lines: plain performance.

図 5 に示すポピュラー歌手による歌声では、ビブラートによるピークは顕著ではない。また、全体にやや 100 Hz 付近の変調周波数の変調のレベルが高くなっているものの、クラシック歌手の特性と大きく異なってはいない。男性歌手の 40 Hz 付近にピークのある特性は、この歌手特有のもののようである。

4. 演奏における歌唱音声の分析

ここでは、CrestMuse プロジェクトにおいて収録したポピュラー曲の演奏における基本周波数の軌跡と、スペクトル包絡の分析結果について紹介する。曲は、プロジェクトのために用意された『RIDE』である。同一の男性歌手により、同じ曲について、(1) できるだけ表情を込めずに楽譜通りに演奏した版 (plain) と、(2) 自分のスタイルで表情豊かに演奏した版 (expressive) とを収録した。

図 6 に歌唱の一部「戯れ言も辛い」の波形と基本周波数の軌跡を示す。plain な演奏は青、表情豊かな演奏は赤で表示している。両者を比較すると、表情豊かな演奏では基本周波数の軌跡に細かな振動が多く含まれていることが分かる。

図 7 は、この様子を拡大したものである。横軸の数値は、図 6 の数値と対応させている。上の段に基本周波数の軌跡を示し、下の段にその差分信号のパワースペクトルを示している。plain な演奏には認められなかった 70 Hz 付近のピークが、表情豊かな演奏では顕著である。RWC データベースで収録された歌唱音声の演奏状況は、ほぼここでの plain に相当するものと考えることができよう。したがって、この 70 Hz 付近の早い変調は、表情豊かな演奏により

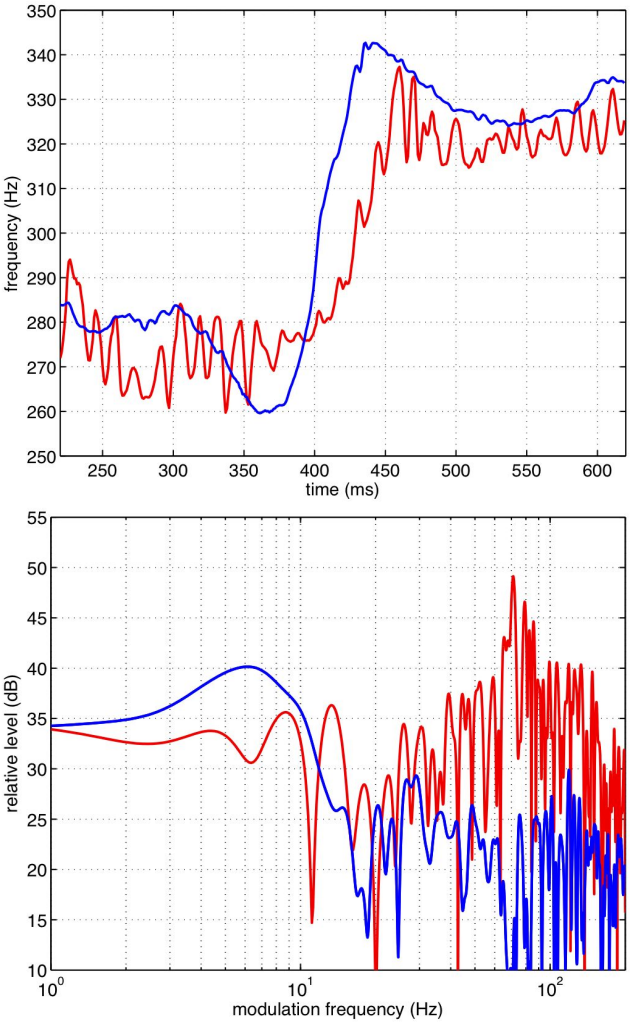


図 7 表現による歌唱音声の基本周波数軌跡と変調周波数のパワースペクトルの変化。上段に波形、下段にパワースペクトルを示す。赤は表情豊かな演奏、青は無表情な演奏を示す。

Fig. 7 F0 trajectories and modulation power spectrum of two singing expressions. Upper plot shows F0 trajectory and lower plot shows modulation frequency power spectrum. Red lines: expressive performance and Blue lines: plain performance.

加えられた特徴と見なして良いであろう。

図 8 に、このときに求められた TANDEM-STRAIGHT によるスペクトル包絡の対応する部分を示す。上に示した表情豊かな演奏には、下の plain な演奏では認められない縦縞上のテクスチャが重なっていることが分かる。このテクスチャの時間方向の周期は、上記の基本周波数軌跡の顕著な変調のピークに対応している。次に、これらの時間的微細構造と知覚との関連を検討する。

4.1 時間的微細構造と知覚

ここでは、時間方向の移動平均を用いて、これらの微細構造を平滑化し、処理したパラメタを用いて歌唱音声を再合成することにより、印象に与える影響を調べた。予備実験では、(1) 元の歌唱音声、(2) パラメタ操作しない再合成

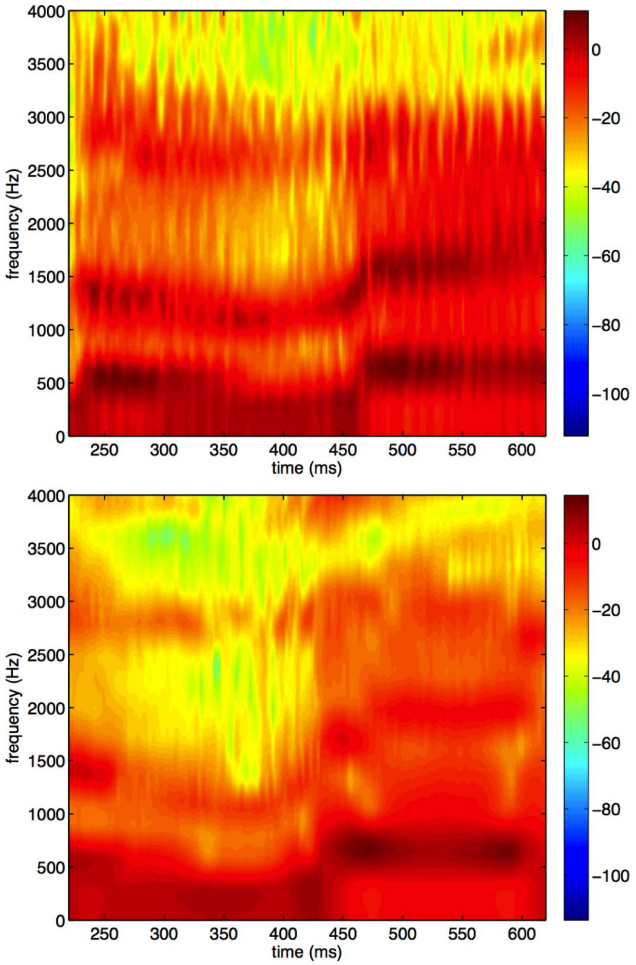


図 8 表現による歌唱音声のスペクトル包絡の変化。上は表情豊かな演奏、下は無表情な演奏を示す。

Fig. 8 Spectral envelope variations due to expressions. Top: expressive performance and Bottom: plain performance.

音声、(3) 基本周波数軌跡を平滑化した再合成音声、(4) スペクトル包絡を時間方向に平滑化した再合成音声、(5) 基本周波数軌跡を平滑化しスペクトル包絡を時間方向に平滑化した再合成音声を用意し、比較試聴した。正式な主観評価実験の結果ではないが、(1) = (2) > (3) > (4) >> (5) の順に、表現の豊かさ（熱く叫んでいる感じ）が失われる印象が得られた。しかし、それらの変化を通じて、歌手の声区や発声の努力の印象には変化が無かったことが興味深い

5. まとめ

新しく提案した高い時間分解能を有する基本周波数分析法を用いて、歌唱音声の基本周波数軌跡の時間的微細構造を調べた。その結果、ポピュラー歌手による表情豊かな演奏において 70 Hz 付近の高速な基本周波数の周波数変調が認められ、併せてスペクトル包絡にも同期した微細構造が認められた。これらの特徴を操作することにより、シャウトなど、これまで困難であった領域の演奏表現を再現し操作する可能性が示された。まだ予備実験の段階であり、これらを組織的検討を進めることが今後の課題である。

謝辞 本研究の一部は、科学研究費 挑戦的萌芽研究による支援を受けた。

参考文献

[1] 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*, Vol. 27, No. 3-4, pp. 187–207 (1999).

[2] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, *ICASSP 2008*, pp. 3933–3936 (2008).

[3] 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. 713–722 (2011).

[4] Ananthapadmanabha, T. and Yegnanarayana, B.: Epoch extraction of voiced speech, *Acoustics, Speech and Signal Processing, IEEE Transactions on*, Vol. 23, No. 6, pp. 562 – 570 (online), DOI: 10.1109/TASSP.1975.1162745 (1975).

[5] Yegnanarayana, B., Murty, S. R. and Rajendran, S.: Analysis of stop consonants in Indian languages using excitation source information in speech signal, *Proc. ISCA ITRW Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark (2008).

[6] Murty, K. S. R. and Yegnanarayana, B.: Epoch Extraction From Speech Signals, *IEEE Trans. ASLP*, Vol. 16, No. 8, pp. 1602–1613 (2008).

[7] Hess, W.: *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag (1983).

[8] 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).

[9] 森勢将雅，河原英紀，西浦信敬：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).

[10] 河原英紀，森勢将雅，西村竜一，入野俊夫：基本波の FM と AM 成分に基づく高速な基本周波数推定法について，日本音響学会聴覚研究会資料， Vol. 41, No. 9, pp. 679–684 (2011).

[11] Kawahara, H., Irino, T. and Morise, M.: An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction, *ICASSP 2011*, pp. 5420–5423 (2011).

[12] de Chevengné, 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).

[13] 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).

[14] 後藤真孝，橋口博樹，西村拓一，岡 隆一：RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌， Vol. 45, No. 3, pp. 728–738 (2004).