Vol.2012-MUS-96 No.7
2012/8/9

歌声・音声分析合成のためのF0適応多重フレーム統合分析に基づくスペクトル包絡と群遅延の推定法

中野 倫靖[1,a)]　後藤 真孝[1,b)]

概要：本稿では、音声（歌声及び話声）の高性能な分析と高品質な合成のために、音声信号からそのスペクトル包絡と群遅延を高い精度と時間分解能で推定する手法を、F0 適応多重フレーム統合分析と名付けて提案する。従来、スペクトル包絡推定に関する研究は数多くなされてきたが、適切な包絡の推定は依然困難な課題である。また群遅延を合成に活用する研究があったが、ピッチマークと呼ばれる時刻情報が必要であった。本研究では、まず、全時刻（全サンプリング点）について、F_0 に適応させた短い時定数の窓を用いて FFT を行い、F_0 適応スペクトルを推定する。次に、分析時刻毎に近傍の複数フレームから F_0 適応スペクトルと群遅延を統合して、最終的なスペクトル包絡と群遅延を得る。スペクトル包絡の推定性能は、14 種類の音サンプル中 13 サンプルにおいて、対数スペクトル距離が 2 種類の既存手法のいずれかよりも低く、8 サンプルにおいて最も低かった。また群遅延を保存して合成できることを確認した。

1. はじめに

ソース・フィルタ分析 [1] は、音声（歌声及び話声）や楽器音を扱う上で重要な信号処理の一つである。観測信号から適切なスペクトル包絡を得ることが出来れば、高性能な分析や高品質な合成、音の変形等の幅広い応用が考えられる。ここで、スペクトル包絡に加えて位相情報まで適切に推定することで、合成音の自然性向上が期待できる。

従来、音の分析においてはスペクトルの振幅情報が重要視されていて、位相情報が考慮されることは少なかった。しかし、音の合成においては、位相が自然性の知覚に重要な役割を果たすことが知られている。例えば、正弦波合成においては、初期位相が自然発話から $\pi/8$ よりも大きくずれると、ずれの大きさに応じて知覚的自然性が単調に減少することが示されている [2]。また、分析合成系では、スペクトル包絡からインパルス応答を求めて単位波形（一周期分の波形）とする際に、最小位相応答が零位相応答よりも自然性が高いことが知られており [3]、自然性向上を目的とした単位波形の位相制御を行う研究 [4] もある。

本研究の目的は、音声や楽器音からスペクトル包絡と位相情報を高い精度と時間分解能で分析し、それを保存したままの高品質な合成を実現することである。その際、ピッチマーク[*1]や音素情報等の付随情報を前提とせず、音の種類の違いによらず安定して分析できるように実装する。

そこで本稿では、F_0 **適応多重フレーム統合分析**と名付けた新しい信号処理手法を提案する。**図 1** にスペクトル包絡と位相情報としての群遅延の推定結果を示す。これ以降、提案手法の実装方法について述べ、スペクトル包絡の推定精度について、正解との対数スペクトル距離を算出して既存手法と比較評価した結果を示す。また群遅延を保存して合成できることを示す。

2. 関連研究

従来、音声信号などの高品質な合成や変形操作のために、信号モデリングに関する数多くの研究がなされてきた。それらの研究では、補足情報を用いない場合、補足情報として F_0 推定を伴う場合、音素ラベルを必要とする場合がある。

2.1 補足情報を用いない信号モデリングに関する研究

代表的な手法として、入力信号を時間周波数平面でのパワースペクトログラムに展開して扱う Phase Vocoder [5,6] がある。周期信号の時間伸縮等が可能だが、非周期性や F_0 の変動等が原因で、品質が劣化してしまう問題がある。

また、古くから知られたスペクトル包絡推定法として、LPC 分析 [7,8] やケプストラム等があり、様々な拡張や組み合わせがなされてきた [9–13]。しかし、包絡概形が LPC

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
[a)] t.nakano [at] aist.go.jp
[b)] m.goto [at] aist.go.jp

[*1] 基本周波数に同期した分析を行う際の、波形の駆動点（かつ分析時刻）を示す時刻情報。声門音源の励起時刻、もしくは基本周期中で振幅が大きい時刻が用いられる [1]。

© 2012 Information Processing Society of Japan

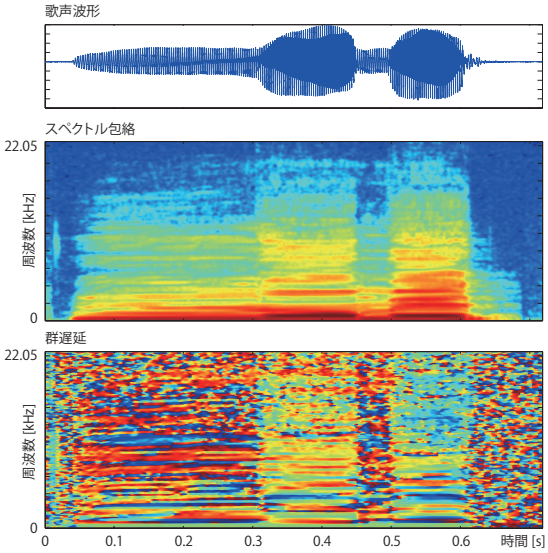

図 1 歌声信号の波形と、そのスペクトル包絡と（正規化された）群遅延。

やケプストラムの分析次数によって決定されるため、次数によっては包絡を適切に表現できない可能性がある。

2.2 補足情報として F_0 推定を伴う分析

時間領域の波形をピッチマークに基づいて単位波形として切り出し、それを基本周期で重畳加算する Pitch Synchronized Overlap-Add (PSOLA) [1,14] が、F_0 に適応した分析として古くから知られている手法である。F_0 の変化にも対応可能であり、位相情報が保存されていることから合成品質が高い。しかし、ピッチマーク付与の難しさや、F_0 の変更や非定常部における品質劣化に関する問題がある。

音声・音楽信号における正弦波モデル [15,16] も、調波構造をモデル化するために F_0 推定を伴う。従来、調波成分と広帯域成分（ノイズ等）のモデル化 [17,18]、スペクトログラムからの推定 [19]、パラメータの反復推定 [20,21]、2 次補間に基づく推定 [22]、時間分解能の向上 [23]、非定常音声での推定 [24,25]、重畳音声での推定 [26] 等の数多くの拡張がなされてきた。これら正弦波モデルの多くは、位相を含めて推定することから高品質な合成が可能であり、高い時間分解能も実現されている [23,24]。

一方、ソースフィルタ分析に基づいたシステム（VOCODER）に、F_0 適応分析の考え方を取り入れた STRAIGHT [27] は、その分析合成品質の高さから世界中の研究コミュニティで使用されている。STRAIGHT では、F_0 適応した平滑化等の処理によって入力音声信号から周期性を除去したスペクトル包絡を得るが、品質の高さに加えて、高い時間分解能も持つ。また、TANDEM 窓によって時間方向の変動を除去する TANDEM-STRAIGHT [28] や、スペクトルピークの強調 [29]、高速計算法 [30] 等への拡張がある。これらの研究では、位相を陽に推定せず、非周期成分[*2]をガウスノイズで畳み込む混合励振による合成方式や、高域の位相（群遅延）を乱数を用いて拡散させる方式、などで合成品質の自然性向上を図っている。しかし、位相の操作に関する基準は明確になっていない。その他、元の音声信号と推定包絡のインパルス応答波形との逆畳込みによって、励起信号を抽出して利用する方法もある [31] が、位相を効率的に表現しているとはいえず、補間や変換操作への応用が困難である。また、群遅延を推定・平滑化して分析合成する研究がある [32,33] が、ピッチマークが必要であった。

以上の研究に加え、スペクトル包絡を混合ガウス分布（GMM）によってモデル化する研究もあり、STRAIGHT スペクトルをモデリングする研究 [34] や、F_0 と包絡の同時最適化による推定を定式化した研究 [35] がある。

これらの研究に共通する問題としては、局所的な観測からの分析である以上、調波構造（F_0 の整数倍の周波数に位置する成分）のみがモデル化され、調波構造間の伝達関数は補間によってしか得られないという問題がある。

2.3 補足情報として音素ラベルを活用する研究

観測できない調波構造間の包絡成分を推定するために、分析時刻と同一の音素で、異なる F_0（異なるフレーム）のスペクトルを統合することで、真の包絡を推定しようとする研究がある [36–38]。単一音のみではなく、音楽音響信号中のボーカルを対象とした研究も存在し [39]、同一の音素であれば、類似した声道形状を持つという仮定に基づく。しかし、正確な音素ラベルが必要であり、また歌声のようにコンテキストの違いによる変動が大きい場合には、過剰な平滑化につながる可能性がある。

3. F_0 適応多重フレーム統合分析

図 2 に、F_0 適応多重フレーム統合分析の概要を示す。本手法では、まず観測信号の全時刻（全サンプリング点）について、F_0 に適応させた短い時定数の窓を用いて FFT 分析を行う（F_0 **適応分析**）。これによって高い時間分解能を持つ F_0 **適応スペクトル**を推定する。次に、分析時刻毎に近傍の複数フレームから F_0 適応スペクトルと群遅延を統合して、最終的なスペクトル包絡と群遅延を得る（**多重フレーム統合分析**）。

図 3 に複数フレームの波形とそれに対応する短時間フーリエ変換（STFT）によるスペクトルと群遅延を示す。それぞれのスペクトルには谷があり、別のフレームではその谷が埋まっているため、これらを統合することで定常なスペクトル包絡が得られる可能性がある。ここで、群遅延のピーク（分析時刻から離れていることを意味する）とスペクトルの谷が対応付いていることから、単一の窓を使った

[*2] 「調波成分の和あるいは周期的パルス列により駆動された応答により記述することのできない成分」と定義されている。

© 2012 Information Processing Society of Japan

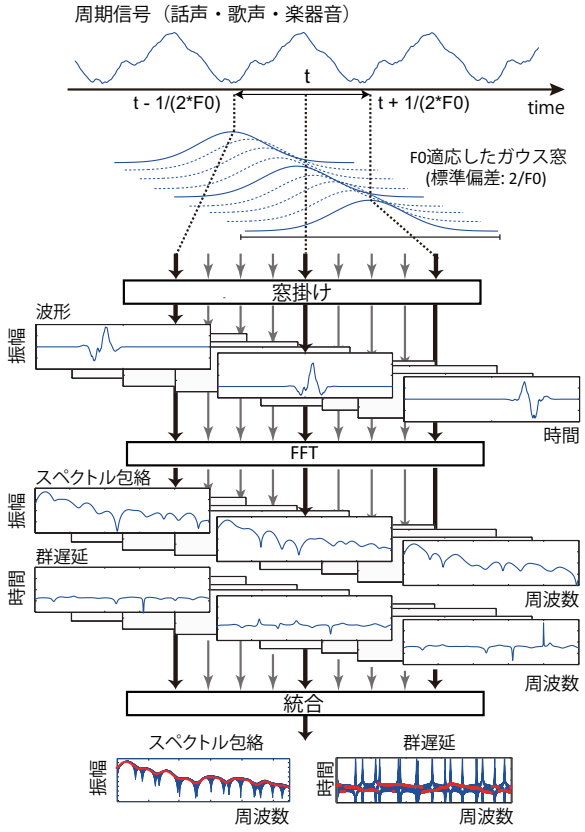

図 2 F0 適応多重フレーム統合分析の概要

だけでは、滑らかな包絡が得られないことが分かる。

推定すべきスペクトル包絡は、この重畳したスペクトルの最大値と最小値の間にあると考え、まず最大値と最小値を計算する。ただし、最大・最小の操作では、時間方向に滑らかな包絡を得られず、F_0 に応じたステップ状の軌跡を描くため、それを平滑化して滑らかにする。最後に、最大包絡と最小包絡の平均として提案スペクトル包絡を得ると同時に、最大から最小の範囲をスペクトル包絡の存在範囲として保存する（**図 4**）。また、推定すべき群遅延としては、最も共振する時刻を表現するために、最大包絡に対応する値を用いる。

提案手法が従来研究（**2.2**, **2.3**）と異なる点は、遠い別の場所ではなく近傍との統合を行う点であり、これによって音素ラベルを必要としない。また、音声波形は、周波数帯域毎に時間方向に少しずつずれて共振している（**3.2** で後述）ため、このような統合処理を行うことで分析時刻によらず（ピッチマークなし）に定常な包絡を推定できる。ただし、観測範囲が局所的である以上、従来手法と同様、完全な調波構造間の観測は行えない。そこで、スペクトル包絡を一つ推定するだけではなくて、その存在範囲を含めて推定することで、応用可能性を拡げることを考える。

3.1 実装条件

F_0 適応分析を行うため、本研究では既に F_0 が何らかの方法によって精度良く推定されていると仮定する。これ以

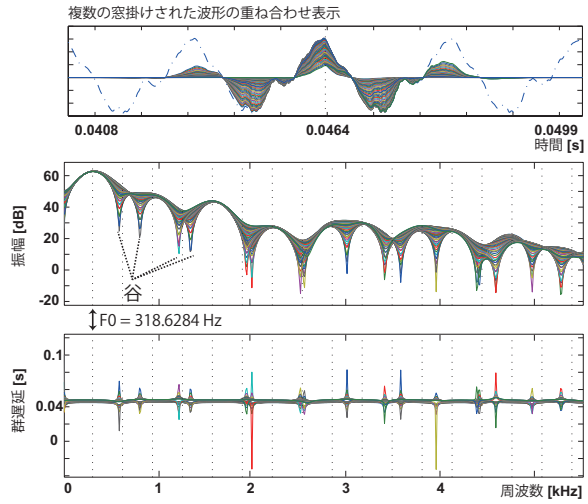

図 3 F_0 に応じた時定数を持つガウス窓を掛けた複数フレームの重畳表示（上図）と、それらに対応するスペクトル（中図）と群遅延（下図）。

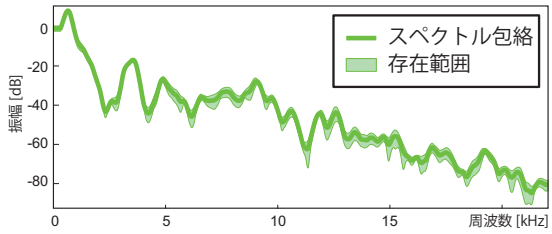

図 4 F_0 適応多重フレーム統合分析によるスペクトル包絡と存在範囲の推定結果。

降、歌声信号はサンプリング周波数 44.1kHz の 16bit モノラル信号を扱い、F_0 適応分析における処理の時間単位は 1/44100 sec、多重フレーム統合分析における処理の時間単位（スペクトル包絡の離散時間）は 1 msec とする。

3.2 F_0 適応分析

本稿では、F_0 適応分析においてガウス窓 $w(\tau)$ を用いる（図 2）。ここで、$\sigma(t)$ は分析時刻 t における基本周波数 $F_0(t)$ によって決まる標準偏差であり、ガウス窓は FFT 長を N として RMS 値で正規化する。

$$w(\tau) = \frac{\hat{w}(\tau)}{\sqrt{(1/N)\sum_{\tau=0}^{N-1}\hat{w}(\tau)^2}} \quad (1)$$

$$\hat{w}(\tau) = \exp(-\frac{\tau^2}{2\sigma(t)^2}) \quad (2)$$

$$\sigma(t) = \frac{1}{F_0(t)} \times \frac{1}{3} \quad (3)$$

ガウス窓の $\sigma(t) = 1/(3 \times F_0(t))$ は分析窓長が基本周期の 2 倍の長さに相当することを意味する（$2 \times 3\sigma(t) = 2/F_0(t)$、図 2）。この窓長は PSOLA 分析などでも用いられ、局所的なスペクトル包絡を近似するための適切な長さであることが知られている [1]。

図 5 に F_0 適応分析の結果例を示す。このようにして得られたスペクトルは、F_0 に起因する時間方向の変動を含

© 2012 Information Processing Society of Japan

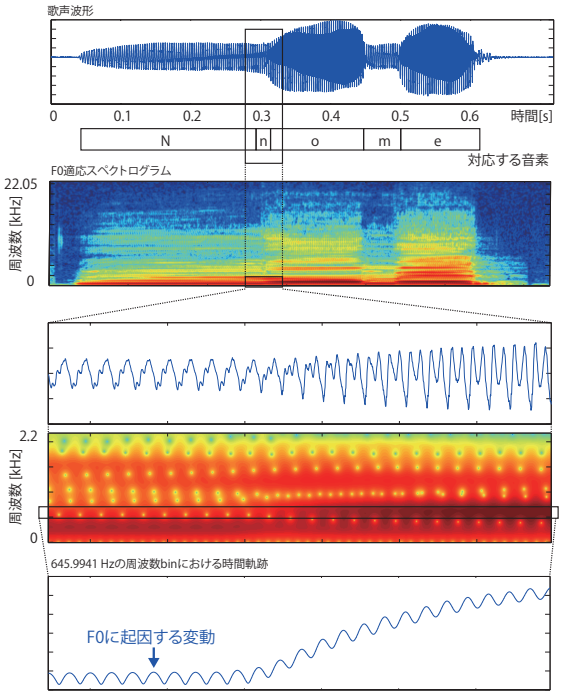

図 5 歌声波形とその F_0 適応スペクトル（上図）とその拡大図（中図）、周波数 645.9961 Hz における時間方向の軌跡（下図）。

み、周波数帯域に応じてピークが時間方向に少しずつずれて出現する。本稿ではこれを F_0 **適応スペクトル**と呼ぶ。

3.3 多重フレーム統合分析

本稿における多重フレーム統合分析では、分析時刻近傍として $-1/(2\times F_0)$〜$1/(2\times F_0)$ の範囲（図 2）の F_0 適応スペクトルを用いる。この範囲は基本周期を意味し、予備実験では、統合範囲をこれ以下とした場合、スペクトルの谷が適切に埋まらなかった。以降、スペクトル包絡と群遅延の推定、それぞれについて述べる。

3.3.1 スペクトル包絡の推定

スペクトル包絡は、統合範囲のスペクトルにおける最大値（最大包絡）と最小値（最小包絡）の平均として定義する。単に最大包絡を用いないのは、分析窓のサイドローブの影響等が含まれている可能性を考慮するためである。ここで、最小包絡には F_0 に起因する多数の谷が残っており、スペクトル包絡として扱いづらい。そこで本稿では、最大包絡を最小包絡にかぶせるように変形することで、包絡概形を保持しながらこれらの谷を除去する。

具体的には、まず最小包絡のピークを算出し、その周波数における最小包絡と最大包絡の振幅の比率を計算する。この変換比率を周波数軸上で線形補間することで、全帯域の変換比率を得る。新しい最小包絡は、最大包絡にこの変換比率を乗じた後、古い最小包絡以上となるように変形して求める。**図 6** にこれらの例と、算出の流れを示す。

また、最大・最小操作によって得られた包絡は、時間方向のステップ状の不連続性が残るため、時間-周波数軸上の

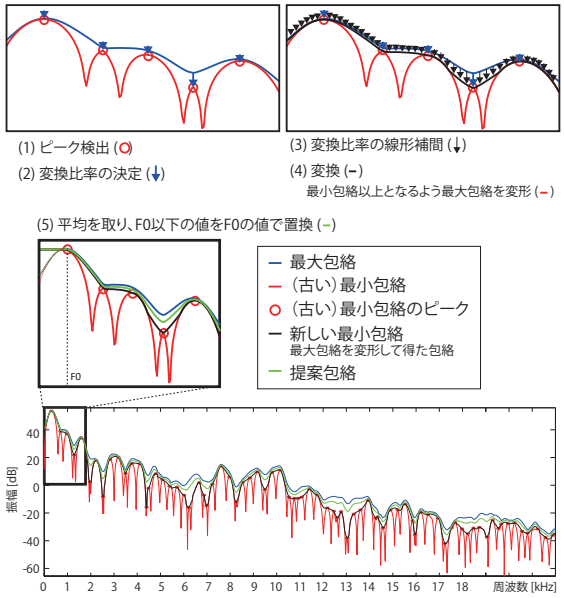

図 6 最大包絡と最小包絡の平均として推定されたスペクトル包絡。

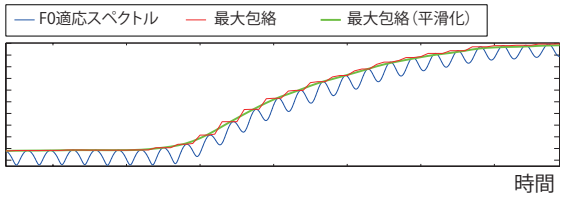

図 7 多重フレーム統合分析によるスペクトルとその 2 次元ローパスフィルタをかけた時間方向の軌跡（図 5 の下図参照）。

2 次元ローパスフィルタによってこれを除去して、時間方向に滑らかなスペクトル包絡を得る（**図 7**）。最後に、F_0 以下の成分が多くの場合に安定して推定できないため、F_0 幅の窓による平滑化に相当する処理として F_0 以下の包絡を F_0 における振幅値で置き換える。

3.3.2 群遅延の推定

群遅延は、統合範囲の中で最も共振する時刻を表現するために、最大包絡に対応する群遅延の値として定義する（**図 8**）。そのようにして求めた群遅延を、推定時刻に対応付けて F_0 適応スペクトル上に重ねて描画した図を**図 9** に示す。この図から分かるように、最大包絡に対応する群遅延は、F_0 適応スペクトルのピーク時刻にほぼ相当する。

このようにして得られた群遅延は、F_0 に対応する基本周期に応じた時間軸方向の広がり（間隔）を持つため、時間軸方向に正規化して扱う。時刻 t、周波数 f における最大包絡に対応する群遅延を $\hat{g}(f,t)$ とすると、基本周期（$1/F_0(t)$）と、$n\times F_0(t)$ に対応する周波数 bin の値 $\hat{g}(f_{n\times F_0(t)},t)$ を用いて、正規化された群遅延 $g(f,t)$ を得る。

$$g(f,t) = \mathrm{mod}\left(\hat{g}(f,t) - \hat{g}(f_{n\times F_0(t)},t), 1/F_0(t)\right) \times \frac{1}{F_0(t)} \quad (4)$$

ここで $\mathrm{mod}\,(x,y)$ は、x を y で割った剰余を意味する。また、$\hat{g}(f,t)-dg$ は、分析時刻の違いにおけるオフセット

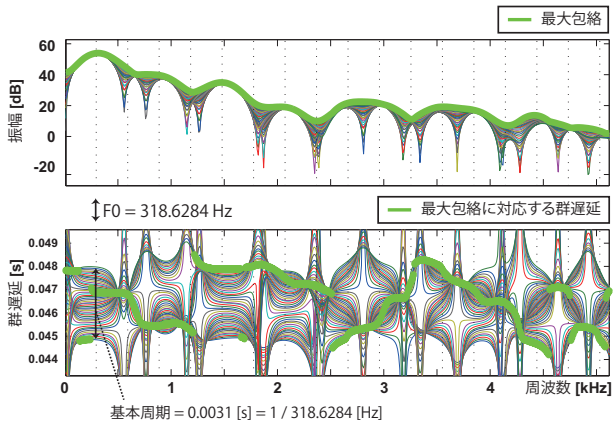

図 8 最大包絡に対応する群遅延。

を除去する操作であり、$n = 1$ もしくは $n = 1.5$ とした[*3]。

以上の操作によって、群遅延 $g(f,t)$ は $(0,1)$ の範囲で正規化された値となる。しかし、基本周期による剰余処理と、基本周期を範囲として統合していることが原因で、次の問題が残る。

(問題 1) 周波数方向に不連続性が発生する。

(問題 2) 時間方向にステップ上の不連続性が発生する。

以下、それぞれの解決法を述べる。

まず**問題 1** は、図 8 の $F_0 = 318.6284$Hz 付近、1.25kHz 付近、1.7kHz 付近等に見られるような不連続性の存在である。この群遅延情報を変形するなど、柔軟に扱いたい場合に、このままでは都合が悪い。そこで、群遅延の値を $(-\pi, \pi)$ の範囲に正規化しなおし、sin と cos で展開すると、この不連続性が連続的に扱える。具体的には、次のように計算する。

$$g_\pi(f,t) = (g(f,t) \times 2\pi) - \pi \tag{5}$$

$$g_x(f,t) = \cos(g_\pi(f,t)) \tag{6}$$

$$g_y(f,t) = \sin(g_\pi(f,t)) \tag{7}$$

続いて**問題 2** は、スペクトル包絡の推定と同様の問題であり、そもそも波形の駆動が基本周期毎に起こることが原因である。ここで、分析合成系として扱うためには、周期間も連続的に変化した値となっていると都合が良いため、$g_x(f,t)$ と $g_y(f,t)$ をそれぞれ平滑化しておく。

最後に、スペクトル包絡同様、F_0 以下の成分が多くの場合に安定して推定できないため、F_0 以下の正規化群遅延を F_0 における値で置き換える[*4]。

3.4 スペクトル包絡と群遅延からの合成

前述のようにして得られたスペクトル包絡と、正規化された群遅延を用いて合成するためには、従来の分析合成システムと同様、時間軸伸縮や振幅の制御を行い、合成の F_0 を指定した上で、それに基づいて単位波形を生成して重畳加算することで合成する。ここで、sin と cos で展開された群遅延 $g_x(f,t)$ と $g_y(f,t)$ から、最終的に以下の計算によって群遅延 $g(f,t)$ に戻してから扱う。

[*3] $n = 1$ 付近では不安定になる場合があり、その場合、調波構造の間の値を基準とした方が、安定した結果を得ることができた。

[*4] 従来研究でも、F_0 以下を零位相で置き換える処理が行われていた [33]。

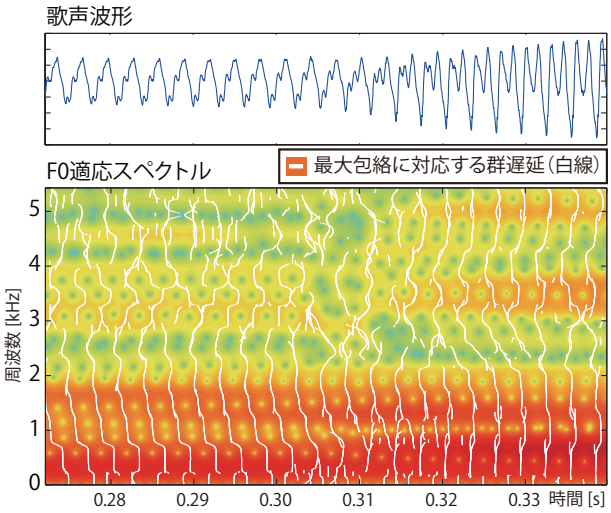

図 9 歌声波形とその F_0 適応スペクトル（図 5 の中図参照）、及び最大包絡に対応する群遅延。

$$g(f,t) = \frac{(g_\pi(f,t) + \pi)}{2\pi} \tag{8}$$

$$g_\pi(f,t) = \begin{cases} \tan^{-1}\left(\frac{g_y(f,t)}{g_x(f,t)}\right) & (g_x(f,t) > 0, g_x(f,t) \neq 0) \\ \tan^{-1}\left(\frac{g_y(f,t)}{g_x(f,t)}\right) + \pi & (g_x(f,t) < 0, g_x(f,t) \neq 0) \\ (3 \times \pi)/2 & (g_r(f,t) < 0, g_x(f,t) = 0) \\ \pi/2 & (g_r(f,t) > 0, g_x(f,t) = 0) \end{cases} \tag{9}$$

ただし、フォルマント周波数が変動する箇所などで、推定された群遅延の形状が急に変わり、特に低域でパワーが大きい場合に合成品質に多大な影響を及ぼすことがある。これは、前述した F_0 に起因する変動（図 5）が、ある周波数帯域において、F_0 以上の速さで変動することが原因と考えられる。例えば図 9 において、500Hz 付近の方が 1500Hz 付近よりも変動が速い。これによって、図中央の前後で、群遅延の形が変わってしまい、単位波形の形も変わる[*5]。現在の実装では、まず、同一の有声区間中では、群遅延 $g(f,t)$ の低域で時間方向の不連続がなるべく発生しないように、新たな共通のオフセットを足して 1 で剰余（正規化されているため）を取った。次に、群遅延の低域に長い時定数のローパスフィルタをかけて、このような瞬間的な変動を除去することで対処した。

そのようにして得られたスペクトル包絡と群遅延を、合成する F_0 の基本周期で取り出し、それぞれの群遅延を合成時の基本周期を係数として乗ずる。その後、群遅延から位相スペクトルに変換し、スペクトル包絡と合わせて単位

[*5] 関係は明らかではないが、正弦波パラメータ推定において、フォルマント周波数と交差する高調波成分が瞬間的に変動する現象がある（ [40] の Fig.4 から読み取れる）。

© 2012 Information Processing Society of Japan

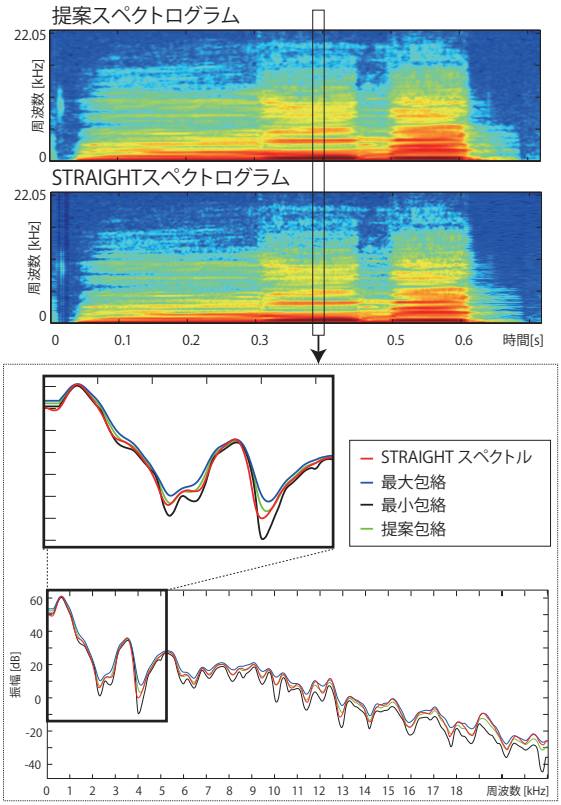

図 10 スペクトログラムの比較。提案手法（上図）、STRAIGHT スペクトログラム（中図）、0.4 秒におけるそれぞれのスペクトル包絡（下図）。

波形を再構成 [32] して、重畳加算する。無声部分に関しては、さらにガウスノイズを畳込む。

ここで、単位波形の配置においては、分析窓としてハニング窓（足して振幅が 1 になる窓）を用いる場合、窓掛けの影響で原音声が変形されることはないが、時間・周波数分解能の向上と、サイドローブの影響（ハニング窓は低次のサイドローブの減衰が少ない）を減らすために、本研究では分析にガウス窓を用いている。そこで、文献 [32] と同様、合成の際に分析時のガウス窓をハニング窓に変換するような窓関数を掛けて合成する[*6]。

4. 実験

提案手法におけるスペクトル包絡の推定精度は、従来、特に性能が高い STRAIGHT [27]、TANDEM-STRAIGHT [28] と比較する。実験には男性の無伴奏歌唱（ソロ）を RWC 研究用音楽データベース [41][*7]から、女性の話声を AIST ハミングデータベース（E008）[42] から、楽器音としてピアノとバイオリンの音を RWC 研究用音楽データベース [41][*8]からそれぞれ用いた。スペクトル包絡の推定精度の比較では、周波数 bin 数を、STRAIGHT で良く用いられる値である 2049 bins（FFT 長が 4096）、分析の時間単位を 1ms とした。提案手法においては、多重フレーム統合分析における統合処理を 1ms ごとに実行する時間単位を意味する。

また、群遅延の推定に関しては、自然音声の分析結果と、群遅延を反映させた合成結果を更に分析した結果を比較する。ここで、群遅延の推定精度を確保するために、スペクトル包絡の推定実験とは異なり、周波数 bin 数を 4097 bins（FFT 長が 8192）と設定して実験した。

[*6] 直流成分が除去されるような波形の違いがあった。ただし、文献 [33] でも述べられているように、聴取印象には影響がほとんどなかった。
[*7] 音楽ジャンル: RWC-MDB-G-2001 No.91。
[*8] 楽器音: ピアノ（RWC-MDB-I-2001, No.01, 011PFNOM）とバイオリン（RWC-MDB-I-2001, No.16, 161VLGLM）。

表 1 実験 B で用いた cascade-type Klatt 合成器 [43] の制御パラメータ。

記号	名称	周波数 (Hz)
F0	基本周波数	125
F1	第 1 フォルマント周波数	250–1250
F2	第 2 フォルマント周波数	750–2250
F3	第 3 フォルマント周波数	2500
F4	第 4 フォルマント周波数	3500
F5	第 5 フォルマント周波数	4500
B1	第 1 フォルマントの帯域幅	62.5
B2	第 2 フォルマントの帯域幅	62.5
B3	第 3 フォルマントの帯域幅	125
B4	第 4 フォルマントの帯域幅	125
B5	第 5 フォルマントの帯域幅	125
FGP	声門共振周波数	0
BGP	声門共振の帯域幅	100

表 2 実験 B における cascade-type Klatt 合成器 [43] の F1 及び F2 の値。

ID	F1 (Hz)	F2 (Hz)	ID	F1 (Hz)	F2 (Hz)
K01	250	750	K04	1000	1500
K02	250	1500	K05	1000	2000
K03	500	1500	K06	500	2000

4.1 実験 A：スペクトル包絡の比較

本実験では、自然音声を対象として STRAIGHT スペクトルと分析結果を比較する。

図 10 に STRAIGHT スペクトログラムと提案スペクトログラムを並べて表示し、0.4 秒におけるスペクトル包絡を重ねて表示している。提案した最大・最小包絡の間に STRAIGHT スペクトルがあり、それは提案スペクトル包絡とほぼ類似していた。さらに、STRAIGHT によって推定した非周期成分を用いて、提案スペクトログラムから音を STRAIGHT で合成した聴取印象は、STRAIGHT スペクトログラムからの再合成と比べて劣るものではなかった。

4.2 実験 B：スペクトル包絡の再現

本実験では、スペクトル包絡と F_0 が既知である合成音を用いて、その推定精度を評価する。具体的には、前述した自然音声及び楽器音を STRAIGHT で分析再合成した音

© 2012 Information Processing Society of Japan

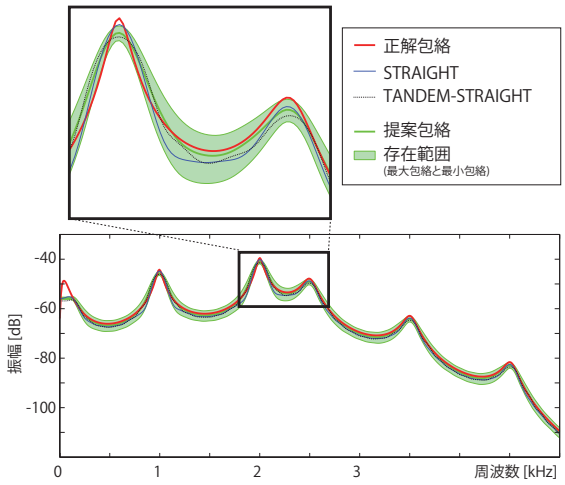

図 11 推定されたスペクトル包絡の比較 (Klatt: K05)。

と、cascade-type Klatt 合成器 [43] によってスペクトル包絡をパラメータ制御した合成音を用いた。

Klatt 合成器に与えたパラメータ一覧を**表 1** に示す。ここで、第 1, 第 2 フォルマント周波数（F1 と F2）の値を、**表 2** に示すように設定してスペクトル包絡を生成し、これらのスペクトル包絡から F_0 を 125 Hz として正弦波を重畳して、6 種類の音を合成した。

推定精度の評価には以下に示す対数スペクトル距離 LSD を用いた。ここで T は有声フレーム数、F は周波数 bin 数（$= F_H - F_L + 1$）、(F_L, F_H) は評価における周波数範囲であり、$S_g(t,f)$ と $S_e(t,f)$ がそれぞれ正解のスペクトル包絡と推定されたスペクトル包絡である。対数スペクトル距離を計算する際には、その形状を評価するために正規化係数 $\alpha(t)$ を $S_g(t,f)$ と $\alpha(t)S_e(t,f)$ の二乗誤差 ϵ^2 が最小になるように算出した。

$$LSD = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{F}\sum_{f=F_L}^{F_H}\left|20\log_{10}\frac{S_g(t,f)}{\alpha(t)\cdot S_e(t,f)}\right| \quad (10)$$

$$\alpha(t) = \frac{\sum_{f=F_L}^{F_H} S_g(t,f)S_e(t,f)}{\sum_{f=F_L}^{F_H} S_e(t,f)^2} \quad (11)$$

$$\epsilon^2 = \sum_{f=F_L}^{F_H}(S_g(t,f) - \alpha(t)S_e(t,f)) \quad (12)$$

表 3 に評価結果を、**図 11** に推定の一例を示す。提案手法によって推定されたスペクトル包絡の対数スペクトル距離は、14 サンプル中 13 サンプルにおいて STRAIGHT と TANDEM-STRAIGHT のいずれかよりも低く、どちらよりも低かったのは 8 サンプルで最も多かった。この結果から、提案手法は高品質な合成と高精度な分析に活用できる可能性が示唆された。

4.3 実験 C：群遅延の再現

男性の無伴奏歌唱を入力として、本手法によってスペクトル包絡と群遅延を推定し、それを再合成した結果を**図 12** に示す。再合成音における群遅延では、低域や全体にかけたローパスフィルタの結果が見られるが、全体的に群遅延を再現して合成できており、合成品質も自然であった。

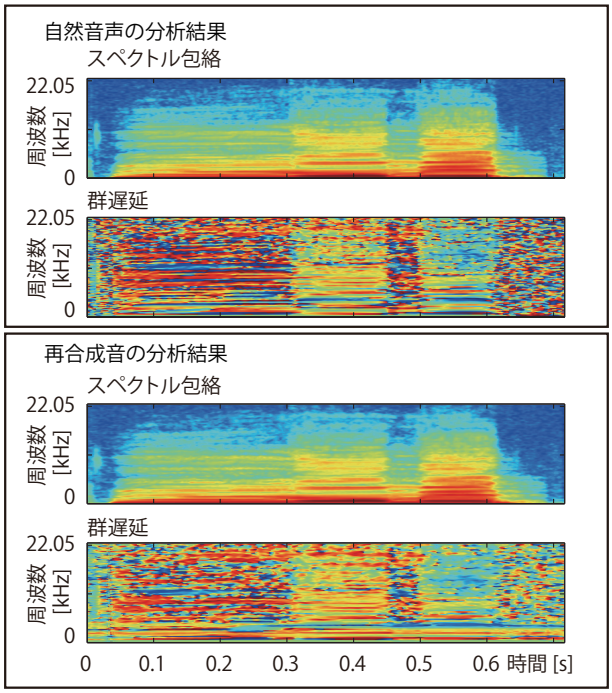

図 12 本手法によって再合成された音の分析結果。

5. おわりに

本論文では、F_0 適応多重フレーム統合分析法を提案し、高い精度と時間分解能でスペクトル包絡と群遅延を推定できることを示した。本手法は、話声・歌声・楽器音を既存手法で分析合成した信号と、Klatt の合成器によって合成した信号を用いて、対数スペクトル距離を算出して評価され、従来手法として STRAIGHT 及び TANDEM-STRAIGHT と比較して、高い精度でスペクトル包絡を分析可能であること示した。さらに、スペクトル包絡は存在可能範囲を同時に推定しており、声質変換やスペクトル形状の変形、素片接続合成等において活用できる可能性がある。

また、群遅延を保存して合成できる可能性も示した。従来の群遅延を用いた研究 [32, 33] では、群遅延を平滑化しても（谷を削っても）合成品質に影響がないことを示したが、それに対して、複数フレームを統合することで谷を適切に埋めることができた。群遅延が周波数帯域毎に、異なる時刻で共振していること（図 9）から、単一のピッチマーキングによる分析を超えて、より詳細に分析できた。

しかし、本稿では合成時にローパスフィルタを掛ける等の処理を行うなど、群遅延を適切に扱いきれていない側面もある。また、現在の群遅延の定義では、音声の特性を完全に表現し切れておらず、今後は改善したい。例えば、最大包絡に対応する群遅延（図 9）では、フォルマント周波数の変動等が原因で、余分なノイズ（誤り）を含んでいる。これは、最大包絡の算出時にピーク検出を行うことで除去

© 2012 Information Processing Society of Japan

表 3 実験 B における各手法で推定されたスペクトル包絡と正解の対数スペクトル距離。最小の値を アンダーライン で、二番目に小さい値を**太字**で示す。

音の種類	長さ [s]	F_L [kHz]	F_H [kHz]	LSD（対数スペクトル距離）[dB] STRAIGHT	TANDEM	Proposed
歌声（男性）	6.5	0	6	1.0981	1.9388	**1.4314**
歌声（男性）	6.5	0	22.05	**2.0682**	2.3215	2.0538
話声（女性）	4.6	0	6	**2.1068**	2.3434	2.0588
話声（女性）	4.6	0	22.05	2.7937	**2.7722**	2.5908
楽器音（ピアノ）	2.9	0	6	3.6600	**3.4127**	3.1232
楽器音（ピアノ）	2.9	0	22.05	4.0024	**3.5951**	3.3649
楽器音（バイオリン）	3.6	0	6	1.1467	1.7994	**1.3794**
楽器音（バイオリン）	3.6	0	22.05	**2.2711**	2.3689	2.1012
Klatt (K01)	0.2	0	5	2.3131	1.6676	**1.9491**
Klatt (K02)	0.2	0	5	3.8462	1.5995	**2.8278**
Klatt (K03)	0.2	0	5	**1.6764**	1.4700	2.2954
Klatt (K04)	0.2	0	5	1.7053	**1.2699**	1.1271
Klatt (K05)	0.2	0	5	1.5759	**1.2353**	1.0643
Klatt (K06)	0.2	0	5	1.1712	**1.2662**	1.8197

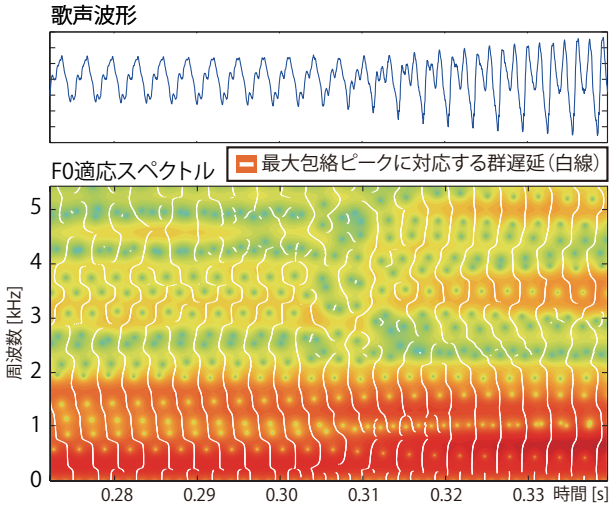

図 13 歌声波形とその F_0 適応スペクトル、及び最大包絡のピークに対応する群遅延。図 9 と比べて群遅延の誤りが減った。

できるため（**図 13**）、包絡推定を含めて改良の余地がある。また、群遅延を適切に扱えれば、有声無声の区別なく扱えるはずなので [32, 33]、現在無声部分の合成でガウスノイズを畳込む必要がある点も今後は解決したい。

さらに、機械学習や音素ラベル情報等の追加情報を活用することで、より精度の高い手法への拡張を検討するとともに、本手法を用いて音声の特性を明らかにしてゆきたい。

謝辞 本研究の一部は、科学技術振興機構 OngaCREST プロジェクトによる支援を受けた。また本研究では、RWC 研究用音楽データベース（音楽ジャンル、楽器音）、AIST ハミングデータベースを使用した。

参 考 文 献

[1] Zölzer, U. and Amatriain, X.: *DAFX - Digital Audio Effects*, Wiley (2002).

[2] 伊藤 仁，矢野雅文：話速変換音声の知覚的自然性に関する検討，電子情報通信学会技術研究報告 EA，pp. 13–18 (2008).

[3] 松原貴司，森勢将雅，西浦敬信：高品質音声合成における有声音の位相特性が知覚に与える影響，日本音響学会聴覚研究会資料，Vol. 40, No. 8, pp. 653–658 (2010).

[4] 濱上知樹：音源波形形状を高調波位相により制御する音声合成方式，日本音響学会誌，Vol. 54, No. 9, pp. 623–631 (1998).

[5] Flanagan, J. and Golden, R.: Phase Vocoder, *Bell System Technical Journal*, Vol. 45, pp. 1493–1509 (1966).

[6] Griffin, D. W.: *Multi-Band Excitation Vocoder*, Technical report (Massachusetts Institute of Technology. Research Laboratory of Electronics) (1987).

[7] Itakura, F. and Saito, S.: Analysis Synthesis Telephony based upon the Maximum Likelihood Method, *Proc. 6th ICA*, pp. C17–20 (1968).

[8] Atal, B. S. and Hanauer, S.: Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, *J. Acoust. Soc. Am.*, Vol. 50, No. 4, pp. 637–655 (1971).

[9] Tokuda, K., Kobayashi, T., Masuko, T. and Imai, S.: Mel-generalized Cepstral Analysis – A Unified Approach to Speech Spectral Estimation, *Proc. ICSLP1994*, pp. 1043–1045 (1994).

[10] 今井 聖，阿部芳春：改良ケプストラム法によるスペクトル包絡の抽出，電子通信学会論文誌，Vol. J62-A, No. 4, pp. 217–223 (1979).

[11] Röbel, A. and Rodet, X.: Efficient Spectral Envelope Estimation and Its Application to Pitch Shifting and Envelope Preservation, *Proc. DAFx2005*, pp. 30–35 (2005).

[12] Villavicencio, F., Röbel, A. and Rodet, X.: Extending Efficient Spectral Envelope Modeling to Mel-frequency Based Representation, *Proc. ICASSP2008*, pp. 1625–1628 (2008).

[13] Villavicencio, F., Röbel, A. and Rodet, X.: Improving LPC Spectral Envelope Extraction of Voiced Speech by

© 2012 Information Processing Society of Japan

True-Envelope Estimation, *Proc. ICASSP2006*, pp. 869–872 (2006).

[14] 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).

[15] McAulay, R. and T.Quatieri: Speech Analysis/Synthesis Based on A Sinusoidal Representation, *IEEE Trans. ASSP*, Vol. 34, No. 4, pp. 744–755 (1986).

[16] Smith, J. and Serra, X.: PARSHL: An Analysis/Synthesis Program for Non-harmonic Sounds Based on A Sinusoidal Representation, *Proc. ICMC 1987*, pp. 290–297 (1987).

[17] Serra, X. and Smith, J.: Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on A Deterministic Plus Stochastic Decomposition, *Computer Music Journal*, Vol. 14, No. 4, pp. 12–24 (1990).

[18] Stylianou, Y.: *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*.

[19] Depalle, P. and Hélie, T.: Extraction of Spectral Peak Parameters Using a Short-time Fourier Transform Modeling and No Sidelobe Windows, *Proc. WASPAA1997* (1997).

[20] George, E. and Smith, M.: Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to The Analysis and Synthesis of Musical Tones, *Journal of the Audio Engineering Society*, Vol. 40, No. 6, pp. 497–515 (1992).

[21] Pantazis, Y., Rosec, O. and Stylianou, Y.: Iterative Estimation of Sinusoidal Signal Parameters, *IEEE Signal Processing Letters*, Vol. 17, No. 5, pp. 461–464 (2010).

[22] Abe, M. and Smith III, J. O.: Design Criteria for Simple Sinusoidal Parameter Estimation based on Quadratic Interpolation of FFT Magnitude Peaks, *Proc. AES 117th Convention* (2004).

[23] Bonada, J.: Wide-Band Harmonic Sinusoidal Modeling, *Proc. DAFx-08*, pp. 265–272 (2008).

[24] Ito, M. and Yano, M.: Sinusoidal Modeling for Nonstationary Voiced Speech based on a Local Vector Transform, *J. Acoust. Soc. Am.*, Vol. 121, No. 3, pp. 1717–1727 (2007).

[25] Pavlovets, A. and Petrovsky, A.: Robust HNR-based Closed-loop Pitch and Harmonic Parameters Estimation, *Proc. INTERSPEECH2011*, pp. 1981–1984 (2011).

[26] Kameoka, H., Ono, N. and Sagayama, S.: Auxiliary Function Approach to Parameter Estimation of Constrained Sinusoidal Model for Monaural Speech Separation, *Proc. ICASSP 2008*, pp. 29–32 (2008).

[27] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A.: Restructuring Speech Representations Using a Pitch Adaptive Time-frequency Smoothing and an Instantaneous Frequency Based on F0 Extraction: Possible Role of a Repetitive Structure in Sounds, *Speech Communication*, Vol. 27, pp. 187–207 (1999).

[28] 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, *Proc. of ICASSP 2008*, pp. 3933–3936 (2008).

[29] 赤桐隼人，森勢将雅，入野俊夫，河原英紀：スペクトルピークを強調した F0 適応型スペクトル包絡抽出法の最適化と評価，電子情報通信学会論文誌，Vol. J94-A, No. 8, pp. 557–567 (2011).

[30] 森勢将雅，松原貴司，中野皓太，西浦敬信：高品質音声合成を目的とした母音の高速スペクトル包絡推定法，電子情報通信学会論文誌，Vol. J94-D, No. 7, pp. 1079–1087 (2011).

[31] Morise, M.: PLATINUM: A Method to Extract Excitation Signals for Voice Synthesis System, *Acoust. Sci. & Tech.*, Vol. 33, No. 2, pp. 123–125 (2012).

[32] 坂野秀樹，陸　金林，中村　哲，鹿野清宏，河原英紀：時間領域平滑化群遅延を用いた短時間位相の効率的表現方法，電子情報通信学会論文誌，Vol. J84-D-II, No. 4, pp. 621–628 (2001).

[33] 坂野秀樹，陸　金林，中村　哲，鹿野清宏，河原英紀：時間領域平滑化群遅延による位相制御を用いた声質制御方式，電子情報通信学会論文誌，Vol. J83-D-II, No. 11, pp. 2276–2282 (2000).

[34] Zolfaghari, P., Watanabe, S., Nakamura, A. and Katagiri, S.: Modelling of the Speech Spectrum Using Mixture of Gaussians, *Proc. ICASSP 2004*, pp. 553–556 (2004).

[35] Kameoka, H., Ono, N. and Sagayama, S.: Speech Spectrum Modeling for Joint Estimation of Spectral Envelope and Fundamental Frequency, Vol. 18, No. 6, pp. 2502–2505 (2006).

[36] Akamine, M. and Kagoshima, T.: Analytic Generation of Synthesis Units by Closed Loop Training for Totally Speaker Driven Text to Tpeech System (TOS Drive TTS), *Proc. ICSLP1998*, pp. 1927–1930 (1998).

[37] Shiga, Y. and King, S.: Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis, *Proc. EUROSPEECH2003*, pp. 1737–1740 (2003).

[38] Toda, T. and Tokuda, K.: Statistical Approach to Vocal Tract Transfer Function Estimation Based on Factor Analyzed Trajectory HMM, *Proc. ICASSP2008*, pp. 3925–3928 (2008).

[39] Fujihara, H., Goto, M. and Okuno, H. G.: A Novel Framework for Recognizing Phonemes of Singing Voice in Polyphonic Music, *Proc. WASPAA2009*, pp. 17–20 (2009).

[40] Ito, M., Ohara, K., Ito, A. and Yano, M.: Source-filter Separation for Nonstationary Voiced Speech Based on Sinusoidal Representation, *Acoust. Sci. & Tech.*, Vol. 31, No. 2, pp. 181–184 (2010).

[41] 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).

[42] 後藤真孝，西村拓一：AIST ハミングデータベース：歌声研究用音楽データベース，情報処理学会研究報告，2005-MUS-61, pp. 7–12 (2005).

[43] Klatt, D. H.: Software for A Cascade/parallel Formant Synthesizer, *J. Acoust. Soc. Am.*, Vol. 67, pp. 971–995 (1980).

© 2012 Information Processing Society of Japan