音楽情報科学 5－5
（1994．2．4）

Sinusoidal modelの特性分析と音合成への適用

小坂 直敏
NTT 基礎研究所

概要 コンピュータ音楽に用いる音色制御技術を確立するための核となるモデルとして直接的に調波毎の信号を推定できるsinusoidal modelを選び、その基本的物理特性について検討した。パラメータ推定はMcAulayおよびQuatieriのアルゴリズムを用いた。正弦波とcharp信号を対象に、S/N_{seg}と瞬時周波数偏差の二つの物理評価尺度を用いて特性を検討した。その結果、正弦波特性はFFTの次数を増やすことによりS/Nsegの下限が向上し、2048点で33dB以上、また瞬時周波数偏差では3Hz以下であることを明らかにした。また、charp信号に対してはS/N_{seg}はFFTの次数に関係なく低下していくが、瞬時周波数偏差では向上していき、音質がさほど劣化しないことを明らかにした。これらの結果から、位相の推定精度はまだ向上する課題が残されているが、音色制御のための基本的ツールとして十分な精度であることが示された。

Analysis of physical characteristics for sinusoidal model and its application to sound synthesis

Naotoshi Osaka
NTT Basic Research Laboratories
3-9-11 Midoricho Musashino-shi, Tokyo, 180

Abstract Physical characteristics of sinusoidal model, which can directly estimate each haromonics, are studied as a central model for timbre control technology for computer music. McAulay and Quatieri algorithm is used for parameter estimation. Both sine waves and charp signals are tested using S/N_{seg} and instantaneous frequency deviation. As a result, it was shown that as the number of FFT increases, sound noise/distortion decreases, and that the model has a sufficient quality as a basic tool for timbre control.

1 はじめに

音合成の研究は音楽情報処理研究でも重要な一分野であり、音声情報処理研究と関連して長い歴史がある。筆者は信号モデルによる音合成研究の枠組の中で、コンピュータ音楽のための音色制御を目的とし、究極は音質の連続的制御が簡易にできるようなモデルの研究を行なっている。この検討を進める上で、特に以下の条件を満たすことを念頭においている。

1. 実在する単音、音声をできるだけ正確に表現できること。
2. 実在しない音をモデルから合成するときの自然性が高いこと。

1. は分析・合成方式に立脚するモデルであること、2. を満たすためには、これらを改良する際に容易なモデル構成が好ましい。そのためにはFM方式のようなモデルパラメータと知覚的な音の単位（ストリーム）とが対応しないモデルは不都合で、ストリームと対応するようパラメータ制御が行なえる構成が便利である。

こうした条件でモデルを構成するための核となるモデルとしてsinusoidal modelを選んだ。これは、glanular synthesis[2][7], phase vocoder[4][5], LPC[6] などと同様に分析合成方式である。ただし、完全に元の信号を復元できるか否かはモデルパラメータの推定方法により異なる。

また、調波構造を持つ音の各調波信号を抽出、再現でき、合成次のパラメータ制御による調波信号の直接的制御が可能であり 2. の条件を満たす。

このモデルは、一般的な楽音表現の代表的手法のひとつで、その音質の良さから近年も盛んに用いられている。しかし、わが国では、音声研究を含めてこれを用いている例はきかない。また、この方式が非常に品質が良いわりには基本的物理特性が報告されていない。そこで、以下では、今後の音色モデルを検討していく上の基礎資料としてMcAulayおよびQuatieri[3]の方法に基づいてモデルを実現し、その物理特性について報告する。

2 Sinusoidal model

2.1 基本式

sinusoidal modelは加算合成方式の範疇である。Moorerが初期モデルの概要を[1]に記している。モデルの基本式は以下に表される。

$$x(n) = \sum_{l=1}^{L} A_l(n)\cos(\theta_l(n)) \qquad (1)$$

ここに、$x(n)$ は n 番目のサンプルの信号, l は調波番号, L は調波の数である。$\theta_l(n)$ は n 番目のサンプルの調波 l の瞬時位相、また、$A_l(n)$ は調波 l の瞬時振幅を表す。L は分析フレームにより異なる値をとる。この定式化は非常に直観的でわかりやすい。しかし、このパラメータの実音からの推定はFlanaganのphase vocoderに影響を受けたMooreのアルゴリズム[1]以来いくつか提案された。これらの中で最近の代表的なパラメータ推定手法がMcAulayおよびQuatieriによるアルゴリズム(MQアルゴリズム)[3]である。

[3]では分析フレーム毎にFFTから得られるローカルピークを算出する。これらピークから得られる複素表現のスペクトルが(1)式の振幅と位相の波形誤差最小とする近似式であることを数理的に示している。さらに、これらの周波数のトラジェクトリをやや経験的な手法でみつけ、時間-周波数軸上で調波の軌跡を決定する。

合成方法は、1)フレーム毎に分析から得られたパラメータで(1)式の表現をそのまま用いてフレーム内で正弦波合成し、さらに、隣あうフレームでオーバーラップアッドをする手法、と2)次式のように瞬時位相関数を3次式で表現するものとがある。

$$\tilde{\theta}_l(n) = \zeta + \gamma t + \alpha t^2 + \beta t^3 \qquad (2)$$

ここでは、今後のパラメータ操作を考える上で、1)瞬時周波数の把握と制御が容易である点、2)位相の把握と制御が容易である点から2)番目の3次式による瞬時位相関数による合成手法を選んだ。なお、この形式のままでは制御パラメータが多過ぎるため、調波の直接的制御はできない。

3 正弦波特性

まず基本的な入出力特性として、正弦波特性を調べた。分析は 10kHz サンプリング、フレーム更新周期 10msec、固定長のハミング窓を用いた。

3.1 周波数 -S/N_{seg} 特性

図 1 に FFT のポイント数を変えた場合の周波数 -S/N_{seg} 特性を示した。S/N_{seg} はフレーム毎に S/N を算出し、その dB 値を全分析フレームで平均したものをいい、時間的にローカルな歪みも数字に反映される。この図より FFT の次数を増やすと特性が向上していくことがわかる。これは次数の増加により、周波数上の補間が正確に行なえ、より正確な周波数推定ができるためである。FFT の離散周波数点と同一の周波数を持つ正弦波の特性はよく、それらの中間の周波数を持つ正弦波の S/N が悪い。この傾向は FFT の次数を増しても変わらないが特性の下限は向上していく。図 2 は帯域を広くしたときの同特性を示している。これも図 1 と同様な特性を示すがサンプリング周波数の半分よりやや低い周波数で不安定な特性となる。 512 ポイントの場合のみ特性の変動が激しいのは、データ収集のための周波数のきざみ幅を 10Hz としたためである。

FFT の次数を 2048 とすれば安定な領域では全体に 33dB 以上の特性が得られている。

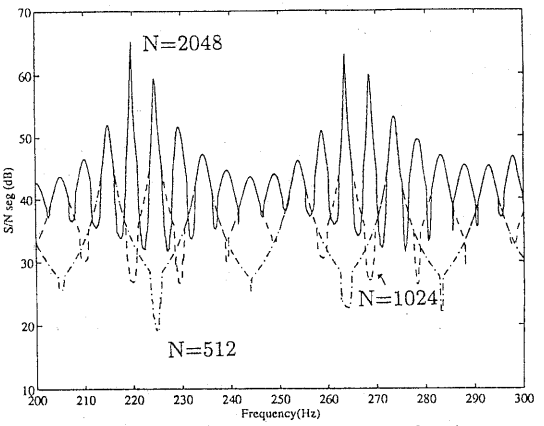

Fig. 1 Frequency-S/Nseg characteristics for sine wave

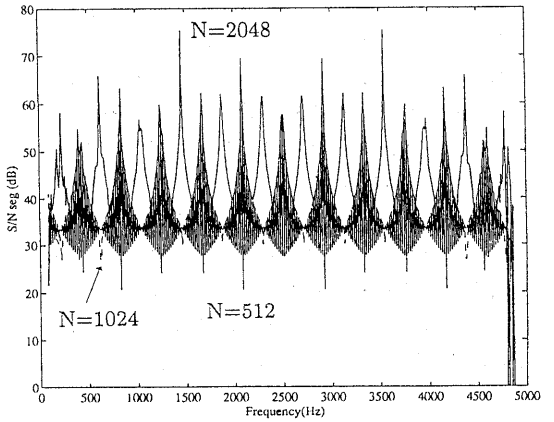

Fig. 2 Frequency-S/Nseg characteristics for sine wave (wide band)

3.2 周波数 - 瞬時周波数偏差特性

S/N は時間領域の評価であるが、周波数領域の評価として、瞬時周波数の特性評価を行なった。評価式は次式による瞬時周波数偏差 d_f を用いた。

$$d_f = \sqrt{\frac{\int_0^T (\tilde{f}(t) - f(t))^2 dt}{T}} \qquad (3)$$

これも S/N_{seg} と同様全フレームで平均をとった値で尺度としている。本方式では瞬時周波数 $f(t)$ は簡単に以下の式で求まる。

$$\tilde{f}(t) = \frac{1}{2\pi}\frac{d}{dt}\tilde{\theta}_l(n) = \frac{1}{2\pi}(\tilde{\omega_0} + 2\alpha t + 3\beta t^2) \quad (4)$$

図 3 に正弦波の周波数 - 瞬時周波数偏差特性を示

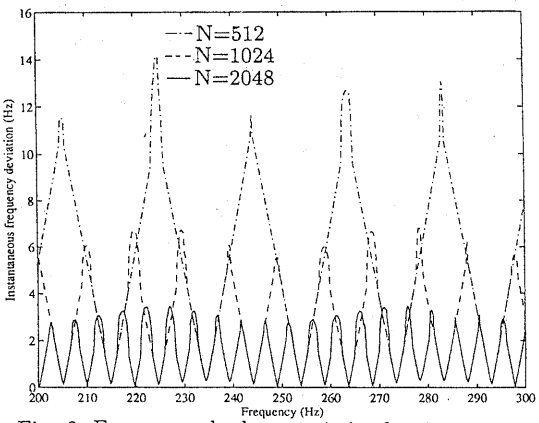

Fig. 3 Frequency-d_f characteristics for sine wave

す。これも図1と同様な傾向を示す。この値の持つ意味について考えてみよう。今基準とする信号との瞬時周波数の誤差が周波数上で分析フレーム毎に $t = 0$ 点で0になるとする。この瞬時周波数誤差 $f_e(t)$ を擬似的に余弦波で表すと、（知覚的には正弦波も余弦波も同等と考えられるため、ここでは余弦波とした。）

$$f_e(t) = f_e(0) + \sqrt{2}d_f \cos(2\pi \frac{1}{T}t) \quad (5)$$

$$\theta_e(t) = 2\pi f_e(0)t + \sqrt{2}d_f T \sin(2\pi \frac{1}{T}t) \quad (6)$$

とかける。ここに、$\theta_e(t)$ は誤差信号の瞬時位相関数、T はフレーム更新周期である。これはFM式

$$y(t) = \sin(\omega_c t + I\sin(\omega_m t)) \quad (7)$$

において $w_c = 2\pi f_e(0), w_m = 2\pi\frac{1}{T}, I = \sqrt{2}d_f T$ としたものである。そこで、ω_c, ω_m, I をパラメータとしたときの $y(t)$ の検知限特性を調べることにより、f_s の値の音質劣化との対応がつけられる。現在はこの実験が行なわれていないため、音質との関係をこれ以上論ずることはできない。

ここで評価値の関連を見るために S/N_{seg} と d_f の両評価値を比較したものを図4に示す。これより、正弦波については、両評価値は比較的対応のよい評価尺度であることがわかる。

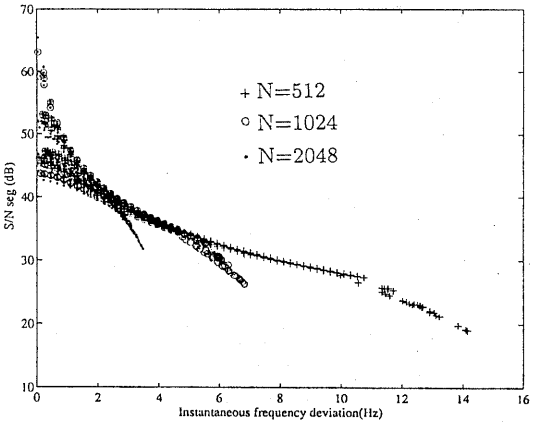

Fig. 4 Correlation between d_f and S/N_{seg}

4 charp信号特性

charp信号は瞬時周波数がtの一次関数となるものをいう。周波数が変化する信号の基本的なものと考えられる。ここでは、1フレームあたり512点のFFTの離散周波数点をいくつ横切るかを周波数変化率 c とし、これをパラメータとした。基準charp信号は次式で表される。

$$f(t) = f_0 + c\frac{f_s/N}{T}t \quad (8)$$

$$\theta(t) = 2\pi f_0 t + c\frac{f_s/N}{2T}t^2 \quad (9)$$

ここに、f_0 は $t = 0$ のときの瞬時周波数、f_s はサンプリング周波数、また $N = 512, T = 0.01$(sec) である。図5にいくつかのcharp信号の瞬時周波数を示す。この分析では $c = 1$ のとき、約1953(Hz/sec)の変化を表す。

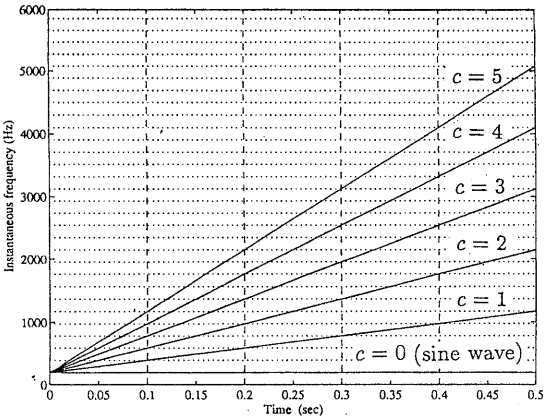

Fig. 5 Instantaneous frequencies of charp signals

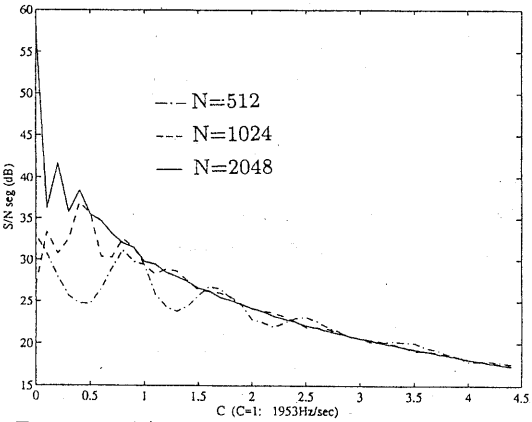

Fig. 6 $c - S/N_{seg}$ characteristics for charp signals

4.1 周波数変化率 -S/N_{seg} 特性

図6にFFTの次数を変えたときの周波数 - S/N_{seg} 特性を示す。この図からわかるとおり、全体にcが大きくなると、S/Nが減少する。cが0に近い場合は正弦波の特性と類似しており、FFTの次数が高い方がS/Nは大きいが、cが大きい場合にFFTの次数が異なっても特性は変わらない。

4.2 周波数変化率 - 瞬時周波数偏差特性

図7に周波数変化率 - 瞬時周波数偏差特性を示す。これより、変化率による変動があるものの、偏差の下限はFFTの次数が向上することにより低下し、歪が低減することがわかる。図8に同図のcが0付近を拡大したものを記した。この図より、c=0から離れると急速に歪が小さくなり、変化する信号の中では正弦波が特異な特性であることがわかる。これは自然な音・音声を分析する場合は非常に都合がよい。

4.3 S/N_{seg} と瞬時周波数偏差特性の相関

ここでも正弦波の場合と同様、d_f とS/N_{seg} を比較したしたものを図9に示す。同図より、charp信号についてはF_sの値に対するS/N_{seg}の弁別があまり良くないことがわかる。すなわち、S/N_{seg}があまり良くなくとも、瞬時周波数偏差小さければ、遅延と同等になるため、音質劣化はさほどないことが想定される。なお、このS/N_{seg}が劣化は図10に示すように、FFTから求めた位相値とt=0におけるcharp信号の偏角（瞬時位相$\theta(t)_{t=0}$の値）が周波数変化率cが大きくなるにつれて真値（図では$\theta(0)$=0としている。）からずれるためである。なお、現在この補正は行なっていない。

5 音合成への適用

これまでの分析でかなりよい音質が期待できそうなので、実際の信号に適用して分析した音の再合成を行なった。1) 楽音として木管の単音、2) 音声として、女声の母音、3) 自然音として水滴の

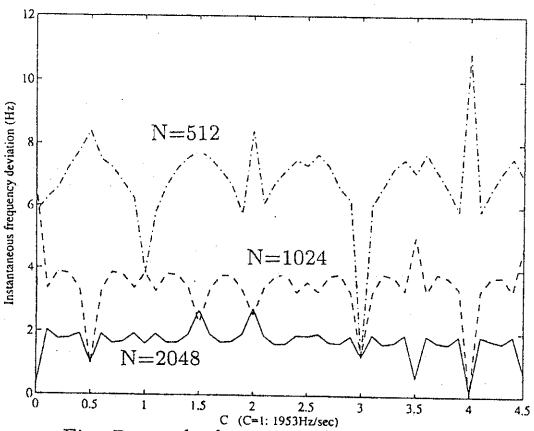

Fig. 7 $c - d_f$ characteristics for charp signals

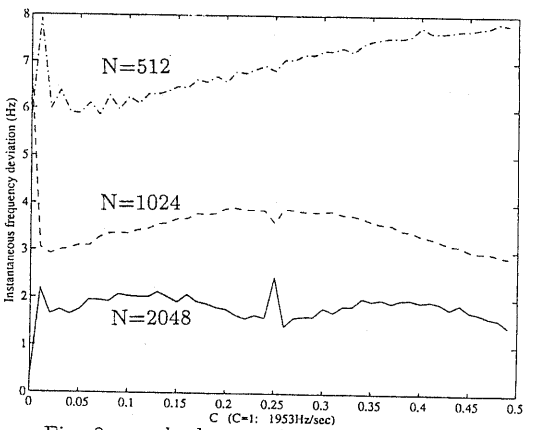

Fig. 8 $c - d_f$ characteristics for charp signals (narrow range)

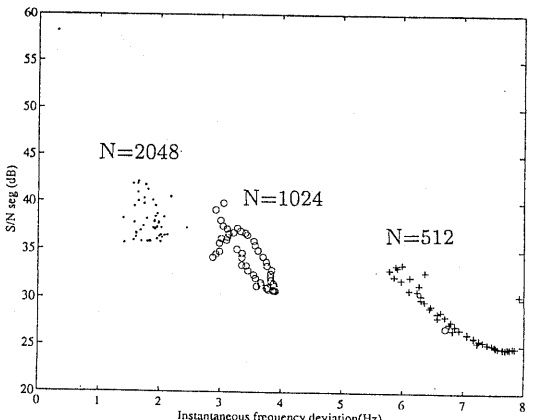

Fig. 9 Correlation between d_f and S/N_{seg} for charp signal

音、などに適用した。この結果、原音とほとんど同等の音質の合成音であった。

6 あとがき

コンピュータ音楽システムのための音色合成技術として重要と考えられる調波信号制御のツールとしてMCアルゴリズムによるsinusoidal modelを選び、基本的な物理特性を分析した。正弦波特性はFFTの次数を増やすことによりS/Nsegの下限が向上し、2048点で33dB以上、また瞬時周波数偏差では3Hz以下であることを明らかにした。また、charp信号に対してはS/N_{seg}はFFTの次数に関係なく低下していくが、瞬時周波数偏差では向上していき、音質が劣化しないことを明らかにした。この他、振幅変化、非定常音、擬似ハーモニクスに対する特性など、その他の諸特性も段階を踏んで特性を調べることが必要だが、今回調べた範囲では十分な精度が得られており、今後の音色合成などのツールとして期待できる。今回の分析により瞬時周波数変化の大きい場合の位相推定精度は改善する余地がある。MCアルゴリズムではFFTで得られた位相は合成時も変化させていないがここも含めて検討する余地がある。

今後は、原音から推定されたパラメータをもとに、音質の劣化を押えながら調波の操作を行ない、以下の課題に取り組んでいく。

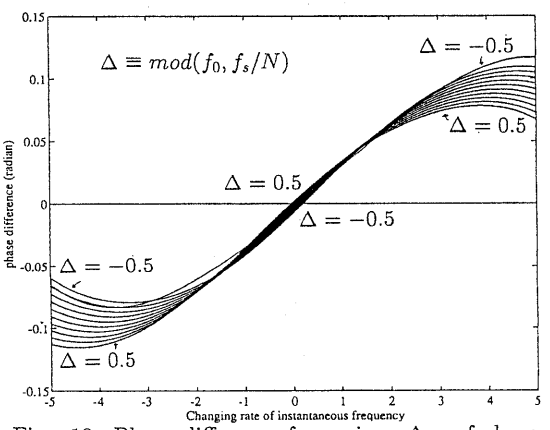

Fig. 10 Phase difference for various Δ s of charp signal

1. フレーム別3次式係数から多フレーム間3次式表現による調波表現の簡略化
2. 調波の伸長、接続
3. 調波の振幅、周波数変化など、調波そのものの任意な特性への連続的制御

これらの過程を経て、音色合成を進めていく予定である。

謝辞 日頃討論していただく基礎研誉田Gの諸氏に感謝します。

参考文献

[1] James A. Moorer,"Signal Processing Aspects of Computer Music: A Survey," *Proceedings of the IEEE,* vol. 65 No. 8, pp. 1108-1137, Aug. 1977.

[2] Curtis Roads, John Strawn,"Granular Synthesis of Sound," *Foundations of Computer Music,* Cambridge, Massachusetts, 1987.

[3] Robert J. McAulay, and Thomas F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoust., Speech,and Signal Processing,*vol. ASSP-34, No. 4, Aug. 1986.

[4] J. L. Flanagan and R. M. Golden,"Phase Vocoder," *Bell Syst. Tech. J.,* Vol. 45, pp. 1493-1509, Nov. 1966.

[5] James A. Moorer,"The Use of the Phase Vocoder in Computer Music Applications," *Journal of the Audio Engineering Society,* vol. 26, No. 1/2 , Jan./Feb. 1978.

[6] J. D. Markel and A. H. Gray, Jr.,"Linear Prediction of Speech," in *Communication and Cybernetics 12,* Spring-Verlag, 1976.

[7] J. S. Lienard,"Speech analysis and reconstruction using short-time elementary waveforms," *IEEE ICASSP'87,* pp. 948-951, 1987.