2 音楽音響信号の実時間加工技術

小野 順貴
東京大学

本稿では，音楽音響信号の加工技術として，我々の研究室で開発してきた，1）スペクトログラムの異方性を用いた調波音・打楽器音分離，2）スペクトログラム伸縮に基づく音高／速度変換を中心に，関連研究を概観しながら紹介する．これらの手法の基礎となる短時間 Fourier 変換とその逆変換，また，実時間実装を可能にするスライディングブロック分析についても論じる．

本稿の目的

音楽信号は分析や認識の対象となるばかりではなく，聞いて楽しむというその本来の性質から，加工もまた重要な研究分野の 1 つである．オーディオアンプのイコライザーやサラウンド再生システムのように，音楽を聴く側で，周波数特性や空間的な広がり感を変化させ，ユーザの好みに合わせて音楽を楽しむことは，従来から行われてきた．近年は，計算機の演算能力の向上と信号処理技術の発展に伴い，さらに新しい種類の加工，高音質化，実時間化などが進められている．本稿ではまず，こうした新しい音楽信号加工技術を概観し，その後，我々の研究室で開発している手法として特に，1）調波音・打楽器音分離，2）音響信号の音高／速度変換，について紹介する．

近年の音楽信号加工技術

■ イコライザー

イコライザーは基本的な音楽信号加工技術の 1 つであり，帯域ごとの周波数ゲインを変化させることによって，スピーカーの周波数特性を補正する，また逆に原音よりも低域や高域を増強するなど，ユーザの好みの周波数特性で音楽を聞くために用いられる．音楽信号を対象とした場合，周波数特性のみならず，単音ごと，声部ごと，楽器ごとの自由な音量操作は究極のイコライジングの 1 つであるが，一般に音楽信号はさまざまな楽器音が重なり合った多重音であり，かつ単音それぞれが幅広い帯域を持っているため，従来の帯域ゲインを調整するタイプのイコライザーでは，どんなに綿密に調整したとしても，単音ごと，声部ごとの音量を独立に制御することはできなかった．こうしたイコライジングのためには，モノラル，もしくはステレオ信号に混合された多重音を個々の音に分離する必要があるが，これが非常に難しい問題であった．しかし近年は，非負値行列分解（Non-negative Matrix Factorization ; NMF），調波時間構造化クラスタリング（Harmonic Temporal Clustering ; HTC）など，スペクトログラムをモデリングする新しい信号処理の枠組みが開拓され，多重音分離に積極的に応用されている [1), 2)]．また，事前知識として，音楽音響信号に同期した MIDI 情報を利用するアプローチも試みられており [3)]，従来は不可能と思われていた高度なイコライザーの実現に向けて研究が進められている．

■ 打楽器成分の抽出／置換

音楽信号においては，メロディや和声を奏でる旋律楽器音だけでなく，打楽器音もまた，楽曲のリズムやムードに関係する重要な要素である．音楽音響信号から打楽器成分を抽出する方法としては，スペクトログラムを前述の NMF によってまず複数の成分に分解し，その後にパターン認識手法を用いて打楽器成分を抽出する手法 [4)]，入力信号を正弦波モデルにフィッティングし，残差成分を打楽器成分として扱う手法 [5)] などが試みられてきている．我々のアプローチに関しては後述するが，調波音成分と打楽器音成分のスペクトログラムの性質に着目し，事前の学習なしにこれらを分離する点に特徴がある．

打楽器音に着目した具体的な音楽信号加工の研究として，たとえば吉井らの Drumix[6)] では，あらかじめ用意した打楽器音のスペクトルテンプレートを楽曲ごとに適応させて発音時刻を検出することにより，楽曲中のドラ

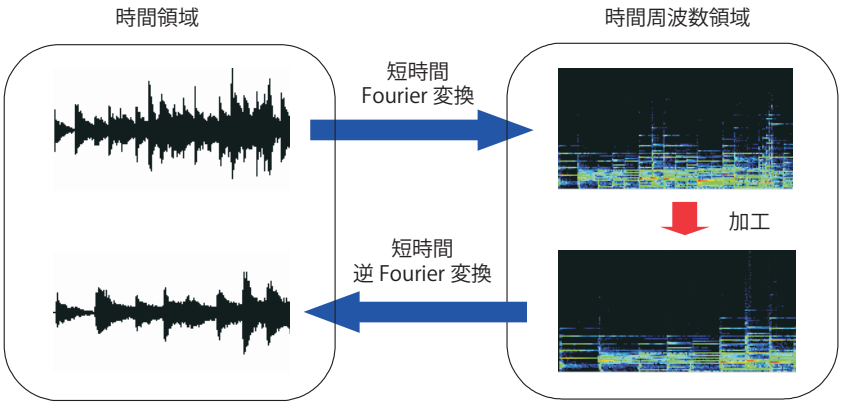

図 -1　短時間 Fourier 変換を用いた信号の加工

ムパートの音量制御だけでなく，音色の置き換えやドラムパターンの編集といった，高度な加工を可能にしている．また角尾らは，楽曲内の打楽器パターンとその構造を楽曲ごとに自動学習し，指定のパターンに置換する手法を開発している [7]．

■ 音高／速度変換

音の高さと速度を自由に変化させる技術もまた，応用の広い加工技術の 1 つである．音声信号の速度を操作する場合には話速変換と呼ばれ，話速を下げることによる聞きやすさの向上，話速を上げることによる速聴のほか，同時通訳支援，外国語学習支援，動画と音声の同期，などに用いられている．音楽の場合においても，音高(キー)や速度の自由な変換は，カラオケや自動伴奏など，従来 MIDI が用いられていたシステムの高音質化につながるほか，ユーザの音楽鑑賞の自由度を大きく広げるものである．我々のアプローチについては後述する．

■ 定位感の操作

ステレオ信号が与える定位感は音楽の臨場感に直結するものであり，その制御については従来からさまざまな研究がなされてきた．イコライザーの場合と同様に，従来は残響感や広がり感といった定位感の全体的な操作が中心であったのに対し，近年は音源分離技術の発展に基づき，音源ごとの定位感の操作のような，より詳細で高度な手法が開発されつつある [8]．

音楽信号加工における短時間 Fourier 変換の利用

音楽信号に含まれるさまざまな音の音高，音長，音色，発音時刻などの情報は，時間波形そのものよりも，時間周波数表現によってよく表現される．そのため，正弦波モデルなどの一部の手法を除き，多くの音楽信号加工技術では，時間波形をまず時間周波数領域に変換し，各成分の増強／低減や伸縮などの加工を行った後に元の時間波形に戻す，といった方法がとられている(図 -1 参照)．一般に信号を時間周波数領域でのエネルギー分布として表現したものはスペクトログラムと呼ばれ，解析が目的の場合には，音階の持つオクターブ構造やピアノロールとの類似性から wavelet 変換や定 Q フィルタバンク分析が好んで用いられる傾向にあるが，加工が目的の場合には高速性，簡便性などの理由から，短時間 Fourier 変換が用いられることが多い．

短時間 Fourier 変換は，音楽に限らず，音声など幅広い対象の分析に古くから用いられている基本的な信号処理法の 1 つであるが，加工に用いるためには逆変換が重要となる．そこではじめに，短時間 Fourier 変換，その逆変換の定義や性質を確認しておく．

まず t を離散時間，$X(t)$ を対象とする離散時系列信号とする．次に短時間 Fourier 変換のパラメータとして，N をフレーム長，R をフレームシフト量とし，それらの比である $Q=N/R$ は簡単のため整数とする．また $W(t)$ を分析窓，$S(t)$ を合成窓とし，これらは $t<0$ または $t \geq N$ では 0 であるものとする．また，$X(t)$ の短時間 Fourier 変換を $H(m, n)$ で表す．ただし，m は時間フレーム番号，n $(0 \leq n \leq N-1)$ は離散周波数である．

短時間 Fourier 変換

分析窓関数による信号の切り出しと離散 Fourier 変換により，式(1)のように定義される．離散 Fourier 変換の時間原点は，フレームごとに切り出された信号の先頭にとられる．

$$H(m,\ n) = \sum_{t=mR}^{mR+N-1} W(t-mR)\, X(t)\, e^{-j2\pi(t-mR)n/N} \tag{1}$$

逆短時間 Fourier 変換

各フレームの離散逆 Fourier 変換に合成窓関数を乗じ，それらをオーバラップして足し合わせる操作として以下

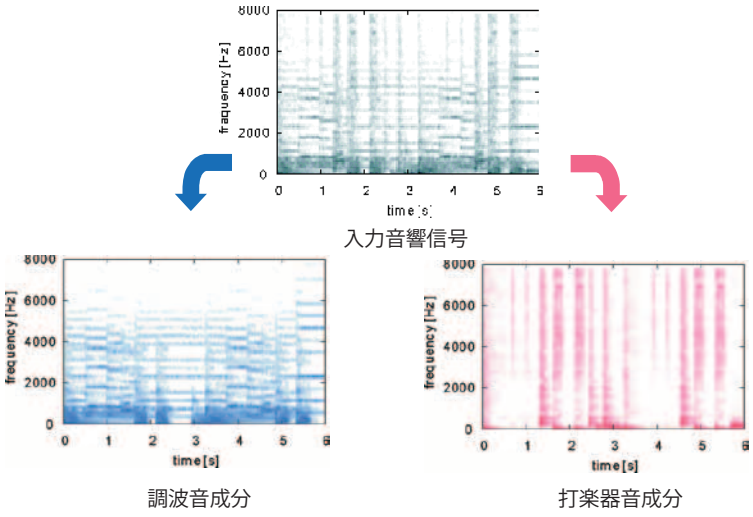

図 -2　調波音・打楽器音分離

のように定義される．

$$X(t) = \sum_m S(t-mR)\left(\frac{1}{N}\sum_{n=0}^{N-1} H(m,\ n)\, e^{j2\pi(t-mR)n/N}\right) \tag{2}$$

各フレームの離散逆 Fourier 変換（式 (2) の括弧内）は，数式上の表現としては周期信号になるため，これを元の分析フレーム部分だけ切り出すことが合成窓の役割である．

完全再構成条件

式 (1) を式 (2) に代入することにより，分析窓と合成窓は以下の条件を満たさなければならないことが分かる．

$$\sum_m W(t-mR)\, S(t-mR) = 1 \tag{3}$$

短時間 Fourier 変換を用いて信号を加工する場合に重要な点としては，下記が挙げられる．

(1) 短時間 Fourier 変換は線形変換である．つまり時間領域での足し算は，短時間 Fourier 領域でも足し算になる．自明ではあるが，これが時間領域で混合している信号の分離を短時間 Fourier 変換領域で考えてよい根拠となっている．

(2) 短時間 Fourier 変換は複素数の表現である．つまり，時間波形に戻すためには振幅と位相の両方の情報が必要である．多くの信号加工手法では，各 $H(m, n)$ の振幅のみを制御し，位相は元の位相がそのまま用いられるが，後述する音高／速度変換などの加工では，陽に位相を推定することが必要になる．

(3) 短時間 Fourier 変換は冗長な表現である．これは，窓関数や離散 Fourier 変換には関係なく，信号をオーバラップして分析することに起因する．つまり，時間領域での 1 サンプル点が Q 個の異なるフレームに含まれ，分析されるので，信号は時間領域の Q 倍の離散点で表現されることになる．このことは，合成窓は式 (3) を満たしさえすればよい，という任意性につながっている．分析窓が Hanning 窓や Hamming 窓の場合には，その中で最も単純な矩形合成窓（$0 \le t \le N-1$ で定数）が用いられることが多い．しかし最小二乗誤差の意味では，下記の式 (4) に示す合成窓が最適であることを指摘しておくのは有用と思われる．詳細な議論は文献 9) を参照のこと．

$$S(t) = W(t) / \sum_m W(t-mR)^2 \tag{4}$$

調波音・打楽器音分離

音楽音響信号を構成する成分は，大きく 2 つの成分に分けることができる．1 つはメロディや和声を奏でる調波音成分，もう 1 つはリズムを担う打楽器成分である．これら 2 つを分離することは，音楽情報検索に関連する多くのタスクにおいて有用な前処理となる．たとえば，多重音解析や和音認識においては，打楽器成分の抑圧は音高推定の手がかりとなる調波構造を強調する効果がある．一方調波音成分の抑圧は，ビート，オンセット，リズムなどの認識をより容易にする．また，分離した 2 つの成分を自由な音量バランスで remix することができれば，新しい音楽イコライザーが実現できることになる．

これを実現するために我々が着目したのは，調波音成分，打楽器音成分のスペクトログラム上での性質の違いである．図 -2 に音楽信号のスペクトログラムの例を示す．図から，明確な縦横の構造を見てとることができる．一般に調波音はメロディ，和音を担うために安定したピッチを持ち，周波数方向には離散的な調波構造をなすことから，スペクトログラム上では横(時間)方向に伸びる線状のパワー分布を示す．一方，打楽器音の波形はイン

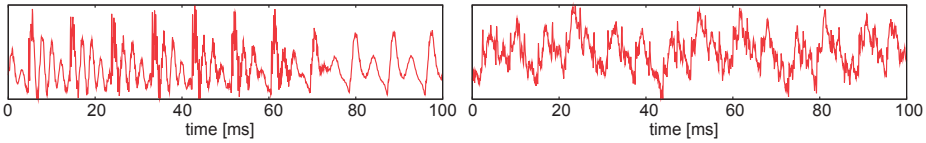

図 -3　音声信号（左）と音楽信号（右）の時間信号波形の例

パルス的で，短い時間区間へのエネルギーの集中と急速な減衰により，スペクトログラム上では縦（周波数）方向に伸びる線状のパワー分布を示す．よって音楽信号のスペクトログラムを縦成分，横成分に分解することによりこれらを分離する，というのが我々の着想である．

実際我々は，まずスペクトログラムを画像と見なし，2 次元フィルタによりこれを分離する手法から検討をはじめた [10]．2 次元フィルタは 2 次元 FFT を用いることにより大変高速に実行でき，ある程度の分離性能が得られることが確認できたが，適切な分離を行うためには，フィルタの遮断周波数等を楽曲ごとにチューニングする必要があった．その後我々は，

(1) 調波音成分のスペクトログラムの横方向の滑らかさ
(2) 打楽器音成分のスペクトログラムの縦方向の滑らかさ
(3) 調波音成分＋打楽器音成分と元のスペクトログラムの近さ

に基づく目的関数を設計し，これを最小化することにより，調波音成分，打楽器音成分を分離する枠組みを開発した [11]．我々はこの一連の手法を，調波音・打楽器音分離（Harmonic/Percussive Sound Separation ; HPSS）と呼んでいる．HPSS は，事前学習が不要であり，短時間 Fourier 変換上の簡単な反復演算で行われ，収束が速い特長がある．

この手法は，後の節で紹介する実時間イコライザとしての応用のほか，単位リズムパターンの抽出と楽曲構造解析，リズムパターンに基づくジャンル認識，打楽器パターンの自動置換，自動和音認識，低音旋律の抽出といった，数多くのタスクの前処理として有効なことが分かっている．特に 2008 年の MIREX（Music Information Retrieval Evaluation eXchange）の和音認識タスクでは，HPSS により調波音を強調したクロマベクトルを特徴量として用いる我々のアルゴリズムが 1 位を獲得した．また，スペクトログラム上の縦横構造が短時間 Fourier 変換のフレーム長に大きく依存することを積極的に利用し，異なるフレーム長の HPSS を組み合わせた多段 HPSS によるボーカル抽出／抑圧，音声強調といった新しい手法と応用が広がりつつある．個々の文献については文献 12) を参照のこと．

スペクトログラム伸縮に基づく音響信号の音高・速度変換

信号の再生速度を変化させると信号波形の周期も変化し，速度とともに音の高さも変わってしまうことから分かるように，音響信号の音高と速度の独立な制御は自明な問題ではない．音声信号の場合にはその周期性に着目し，単位周期波形を切り出して接続したり削除したりする，時間領域での波形接続方式が多く用いられている．しかしながら音楽信号は多くの場合，さまざまな音高（周期）を持つ信号が重なりあう多重音であり，図 -3 に示すように，一般に明確な単位周期波形があるとは限らない．

そもそも音響信号の音高／速度変換は，信号分離などの問題とは異なり，真値や正解が存在するわけではなく，人間にとって自然に聞こえるような音をつくる，というのが 1 つの基準である．人間の聴覚系の知覚においてはスペクトログラムに類似した表現が用いられているという知見に基づくならば，図 -4 のようにスペクトログラムを周波数方向，もしくは時間方向に伸縮し，そのようなスペクトログラムを持つ信号波形を合成すれば，元の信号とほぼ同じ性質を保ちつつ，音高，もしくは速度が異なる信号が得られるのではないか，というのが我々の着眼である．

ここで問題となるのは，伸縮したスペクトログラム，すなわち時間周波数領域の振幅（もしくはエネルギー）分布に対しどのような位相を与えるか，ということである．一般にどんな適当な位相を与えたとしても，式 (2) の公式に従い逆短時間 Fourier 変換を計算することはできる．しかし一般には，そうして得られた時間信号をもう一度短時間 Fourier 変換すると，元には戻らない．すなわち，望んだスペクトログラムに対応する信号波形を得たことにはならない点に注意が必要である．これは前述の通り，短時間 Fourier 変換が冗長な表現であり，フレーム同士は完全に独立ではない，ということが原因である．

図 -5 に簡単な例を示す．左上は，ある音楽信号に 1/2 オーバラップで分析窓をかけて得た隣接フレーム波形である．窓関数の影響はあるものの，細かい凹凸は当然のことながら一致している．これに合成窓をかけ，オーバラップして加算すれば元の信号波形に戻り，また分析窓

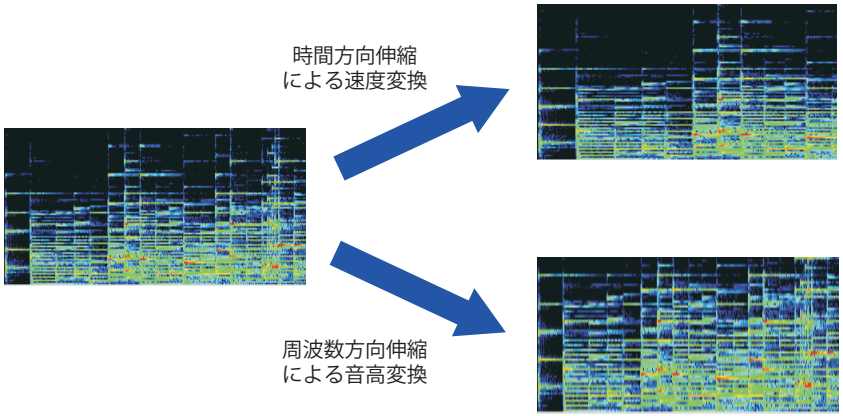

図 -4　スペクトログラムの伸縮に基づく音高・速度変換

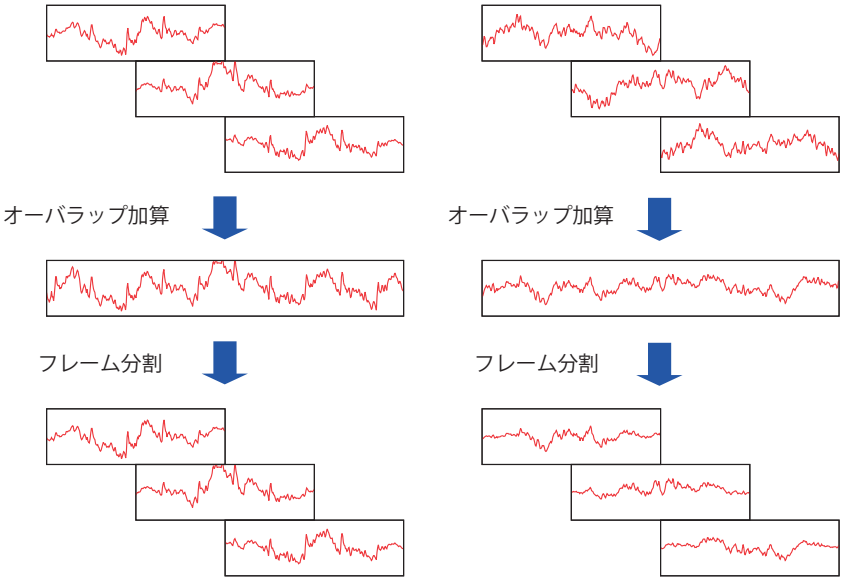

図 -5　整合性のある隣接フレーム（左）と整合性のない隣接フレーム（右）

をかけてフレーム化すれば，同じフレーム波形が得られる．一方右上は，左上の各フレームを一度 Fourier 変換し，適当に位相を変化させて逆 Fourier 変換した隣接フレーム波形である．つまり左上と右上は，スペクトログラムの振幅はまったく等しい．しかし，右上はフレーム間のつじつまが合っておらず，合成窓をかけ，オーバラップ加算して信号波形に戻し，また分析窓をかけてフレーム化すると，元のフレーム波形とは異なるものが得られてしまう．逆にいえば，隣接フレームのつじつまが合うこと，つまり逆短時間 Fourier 変換してまた短時間 Fourier 変換したらできるだけ元に戻るということを，短時間 Fourier 変換の振幅分布から位相分布を決める基準として用いることができる．

与えられたスペクトログラムに対応する信号波形を生成する具体的な解法として，1）設計した振幅スペクトログラム分布に適当な初期位相をつける，2）逆短時間 Fourier 変換，3）短時間 Fourier 変換（これにより，フレーム間のつじつまが合った位相が付加されるが，振幅分布も与えたものから変化してしまう），4）スペクトログラムの振幅を設計したものに置き換える，5）2）に戻る，という反復解法を与えたのが Griffin and Lim[9] である．

我々はこの手法を発展させ，フレームごとのリサンプリングと組み合わせた音高／速度変換，スペクトル包絡を保った音高変換，後述するそれらの実時間化，非線形な時間軸の変換，などに応用し[13]，また位相推定法自体の新しい高速解法の研究も行っている[14]．

スライディングブロック分析による実時間処理

前述の信号加工技術の実時間化は，動的な音楽鑑賞の幅をさらに広げるとともに，他のシステムとの統合などにも大変有用である．一般に実時間処理を行う場合には，1）処理時間が入力データ長よりも短い（高速性），2）未来のデータを使わない（因果性），の 2 つが必要とされる．反復解法であっても収束が速ければ 1）を満たし得

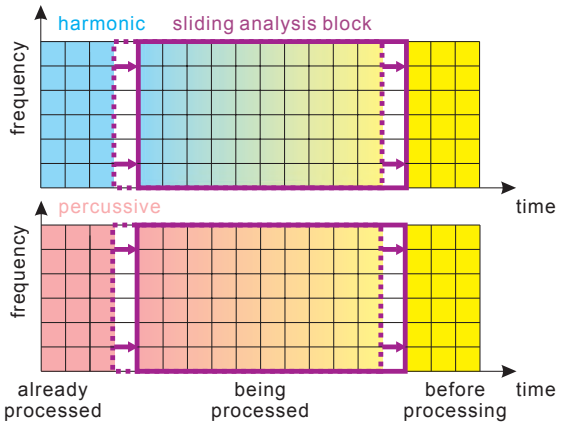

図 -6　スライディングブロック分析

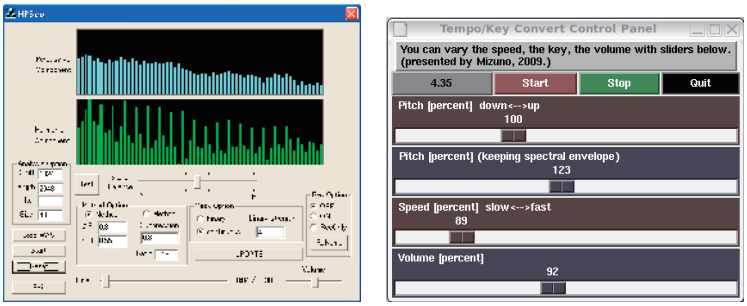

図 -7　HPSS（左）と音高／速度変換（右）の GUI 付き実時間実装

るし，2)に関しても実用上は，完全な因果性を満たさなくても，ある程度の時間遅延は許容されることが多い．

以上を踏まえ，我々は前述の 2 つの手法を実時間化するために，スライディングブロック分析を試みた．スライディングブロック分析とは，反復更新を行う適当なブロックサイズ（たとえば N フレーム）を定め，このブロックを逐次的にずらしていくことにより行う手法である．図 -6 は HPSS の例である．各フレームでは分析ブロックがスライドするごとに 1 回ずつ，最終的には N 回の反復演算が行われる．よって，1）N 回で反復演算がほぼ収束する，2）N フレームのブロック分析に必要な演算時間が，1 フレームシフト以内に収まる，3）N フレーム分の時間遅延が実用上許される，という条件が満たされるならば，この手法により実時間化が可能となる．我々の実時間実装（図 -7）では，HPSS では 1/2 フレームシフトで 60 フレーム，音高／速度変換では 1/8 フレームシフトで 8 フレームを分析ブロックとしている．

今後の展望

近年の音楽信号処理を信号加工の観点から，HPSS と音高／速度変換を中心に紹介した．本文中でも述べたように，両技術とも音楽信号加工が当初の目的であったが，その応用範囲が音声信号など他の音響信号の分析や加工に広がりつつある．音楽信号処理の近年の発展は，検索やリコメンデーションといった大きな需要があることのみならず，音楽信号それ自体が複雑かつ構造的な特徴を持っており，信号処理自体の対象としても面白いことが大きな発展の一因になっているように筆者には思われる．wavelet 解析が地震波の研究から発したように，音楽信号処理という土壌から，他分野へ応用可能な新しい信号処理技術が生まれてくる可能性があり，今後の発展がますます楽しみな研究分野と思われる．

参考文献

1）Smaragdis, P. and Brown, J. C. : Non-Negative Matrix Factorization for Polyphonic Music Transcription, Proc. WASPAA, pp.177-180 (2003).
2）Kameoka, H., Nishimoto, T. and Sagayama, S. : A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering, IEEE Trans. ASLP, Vol.15, No.3, pp.982-994 (Mar. 2007).
3）糸山，後藤，駒谷，尾形，奥乃：楽譜情報を援用した多重奏音楽音響信号の音源分離と調波・非調波統合モデルの制約付パラメータ推定の同時実現，情報処理学会論文誌，Vol.49, No.3, pp.1465-1479 (Mar. 2008).
4）Helen, M. and Virtanen, T. : Separation of Drums from Polyphonic Music Using Non-Negative Matrix Factorization and Support Vector Machine, Proc. EUSIPCO (Sep. 2005).
5）Gillet, O. and Richard, G. : Extraction and Remixing of Drum Tracks from Polyphonic Music Signals, Proc. WASPAA, pp.315-318 (2005).
6）Yoshii, K., Goto, M., Komatani, K., Ogata, T. and Okuno, G. H. : Drumix : An Audio Player with Real-time Drum-part Rearrangement Functions for Active Music Listening, IPSJ Journal, Vol.48, No.4, pp.134-144 (2007).
7）角尾，小野，嵯峨山：音楽音響信号中の打楽器パターンの自動置換，日本音響学会秋季研究発表会講演集，pp.875-876 (Sep. 2008).
8）Haraguchi, Y., Miyabe, S., Saruwatari, H., Shikano, K. and Nomura, T. : Source-Oriented Localization Control of Stereo Audio Signals Based on Blind Source Separation, Proc. ICASSP, pp.177-180 (Apr. 2008).
9）Griffin, D. W. and Lim, J. S. : Signal Estimation from Modified Short-Time Fourier Transform, IEEE Trans. ASSP, Vol.32, No.2, pp.236-243 (Apr. 1984).
10）宮本，立薗，ルルー，亀岡，小野，嵯峨山：スペクトログラム 2 次元フィルタによる調波音・打楽器音の分離，日本音響学会秋季研究発表会講演集，pp.825-826 (Sep. 2007).
11）Ono, N., Miyamoto, K., Kameoka, H. and Sagayama, S. : A Real-time Equalizer of Harmonic and Percussive Components in Music Signals, Proc. ISMIR, pp.139-144 (Sep. 2008).
12）http://hil.t.u-tokyo.ac.jp/publications/publist.php
13）水野，ルルー，小野，嵯峨山：パワースペクトログラムの伸縮と無矛盾位相付加に基づく音楽音響信号の実時間テンポ／ピッチ変換，日本音響学会春季研究発表会講演集，pp.843-844 (Mar. 2009).
14）Le Roux, J., Ono, N. and Sagayama, S. : Explicit Consistency Constraints for STFT Spectrograms and Their Application to Phase Reconstruction, Proc. SAPA (Sep. 2008).

（平成 21 年 7 月 3 日受付）

小野 順貴（正会員）　onono@hil.t.u-tokyo.ac.jp

2001 年東京大学大学院工学系研究科計数工学専攻博士課程修了．博士（工学）．同年より同大学院情報理工学系研究科助手．2005 年より，同講師．計測工学，音響・音楽信号処理，パターン認識の教育・研究に従事．日本音響学会佐藤論文賞，栗屋学術奨励賞，電気学会センサ・マイクロマシンシンポジウム五十嵐賞，ISIE Best Paper Award などを受賞．日本音響学会，電気学会，計測自動制御学会，IEEE 各会員．