補助関数法による音楽音響信号へのMIDI信号の詳細なフィッティング*

☆高宗典玄[1]，亀岡弘和[1,2]，土屋政人[1]，嵯峨山茂樹[1]（[1]東大院・情報理工，[2]NTT CS研）

1 はじめに

本報告では，所与の音楽音響信号をMIDI (Musical Instrument Digital Interface) 信号で音響的に良く再現するよう，MIDI信号中の各単音のベロシティやオンセット時刻などのパラメータを自動推定する方法を提案する．

MIDI信号は，個々のノートの音高，発音タイミング (オンセット時刻) や発音強度 (ベロシティ) などの情報で構成され，楽譜を書いたり楽譜上で音符を操作するような感覚で作編曲や楽曲加工を手軽に行うことが可能である．一方で，表情豊かな人間の演奏らしい演奏をMIDIで打ち込むのは，非常に入念な人手のチューニングを要し，必ずしも容易ではない．しかし，表情豊かな人間らしい演奏を自動でMIDIに打ち込めるようになれば，音楽の加工の可能性が広がるだろう．例えば，表情豊かな人間の演奏らしい演奏の音響信号に対し，その演奏をMIDI信号で再現するようMIDIのパラメータを自動的にチューニングすることが出来れば，自動表情付け [1] の学習データの自動生成や音響信号を加工する際のインデックスとして用いることが期待できる．

MIDIと音楽音響信号の時間整合をとる問題に関してはスコアアライメントと呼ぶ技術 (例えば，[2,3]) が有効であるが，上述のような目的のためには，テンポのような大域的な情報の整合をとるだけでなく各ノート毎の発音時刻，音長，音量などの詳細な情報の整合をいかに正確にとれるかが重要課題となる [4–6]．MIDIと音楽音響信号のテンポ，各ノート毎の発音時刻，音長，音量の整合をとることをスコアアラインメントと区別してオブジェクトアラインメントと呼ぶことにする．オブジェクトアラインメントを目的とした従来の方法は，多重音解析を高精度に行う (音楽音響信号からテンポ，各ノート毎の発音時刻，音長，音量を推定する) ための補助情報としてMIDIを用いる，という考え方のものが多かった．これに対し，本研究では，MIDIパラメータから実際に生成される音響信号が所与の音楽音響信号に音響的に似るようにMIDIパラメータを決定する問題を扱い，本発表ではそのための方法を提案する．

2 音楽音響信号とMIDI信号のスペクトログラムフィッティング

2.1 MIDIパラメータ推定問題の定式化

MIDI信号が観測音響信号を「音響的」に再現するには，MIDI信号から生成される音響信号と観測音響信号の「近さ」を考えるとき，波形よりもむしろスペクトログラムで距離を考えるべきである．そこで，MIDI信号から生成される音響信号のスペクトログラムを考えたい．

MIDI信号から生成される音響信号は，各音がもつMIDIパラメータに従う単音の音響信号の重ね合わせで表現できる．そこで，MIDI信号から生成される音響信号のスペクトログラムを単音の音響信号のスペクトログラムの和で近似できると仮定すると，MIDIパラメータの推定は以下のように観測音響信号とMIDIが生成する音響信号とのスペクトログラム上の乖離度を示す目的関数 $J(\Theta)$

$$J(\Theta) = \sum_{\omega,t} D\left(Y_{\omega,t} \middle| \sum_k \alpha_k X_{k,\omega,t-\tau_k}\right), \quad (1)$$

$$X_{\omega,t,k} = f(\theta_k) \quad (2)$$

を最小化する問題として捉えられる．

ここで，ω と t は周波数と時間方向の添え字であり，k は楽譜上の何番目の音符かを表わす添え字である．また，$Y_{\omega,t}$ は観測音響信号のスペクトログラム，$X_{k,\omega,t}$ は k 番目のノートに対応するMIDIパラメータから生成される音響信号のスペクロトグラムであり，α_k はそのエネルギー，τ_k は発音時刻を表す．よって，推定すべきパラメータは，$\{\alpha_k, \tau_k\}_{1\le k\le K}$ であり，これらをまとめて Θ で表す．D はスペクトログラム間の距離を表し，例えば二乗誤差，Iダイバージェンス，板倉斎藤擬距離

$$D_{\mathrm{EU}}(y|x) = (y-x)^2, \quad (3)$$

$$D_{\mathrm{KL}}(y|x) = y\log\frac{y}{x} - y + x, \quad (4)$$

$$D_{\mathrm{IS}}(y|x) = \frac{y}{x} - \log\frac{y}{x} - 1 \quad (5)$$

* “Optimal MIDI-fitting to music signals with auxiliary function approach” by Takamune Norihiro[1], Kameoka Hirokazu[1,2], Tsuchiya Masato[1], Sagayama Shigeki[1] ([1]Univ. of Tokyo, [2]NTT CS Lab.).

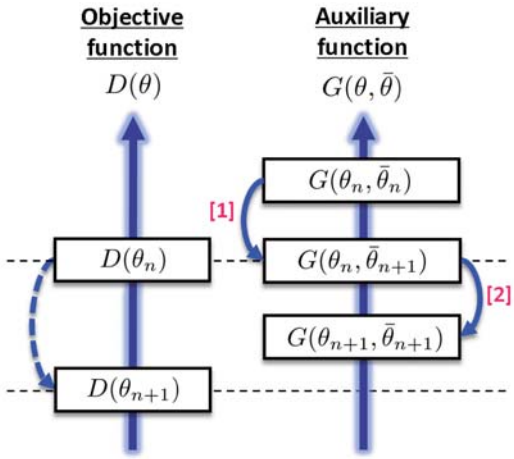

Fig. 1　補助関数法の原理

や，[7] で用いられている

$$D_{\mathrm{KA}}(y|x) \equiv \begin{cases} \beta/\gamma & (|y-x| \le \beta - c) \\ \frac{(y-x-|\beta-c|)^2}{4c\gamma} + \frac{\beta}{\gamma} & (\beta - c < |y-x| \le \beta + c) \\ |y-x|/\gamma & (|y-x| > \beta + c) \end{cases} \tag{6}$$

のような距離関数で測ることができる．

この最適化問題は膨大な解空間を持つため，単純に大域最適解を探すのは困難である．もし，目的関数が単音ごとのスペクトログラム間距離の和に分離した形をしていれば，各単音ごとに最適な MIDI パラメータを探索することができて好都合であるが，残念ながらこの目的関数ではそうはなっていない．そこで，我々は，目的関数の上限となる補助関数を反復的に降下させることで目的関数を間接的に降下していく方法をベースにし，その補助関数として単音ごとのスペクトログラム間距離の和に分離した形をとるものをうまく設計することで，当該最適化問題の解を見通し良く探索することができると考えた．補助関数の反復降下による目的関数の降下方法を補助関数法と呼び，音響信号処理分野で近年様々な最適化問題に適用されている [8,9]．補助関数の定義と補助関数法の原理は以下のとおりである．

定義 1. θ をパラメータとする目的関数 $D(\theta)$ に対し，

$$D(\theta) = \min_{\bar{\theta}} G(\theta, \bar{\theta}) \tag{7}$$

が成り立つとき，$G(\theta,\bar{\theta})$ を $D(\theta)$ の**補助関数** *(Auxiliary function)*，$\bar{\theta}$ を**補助変数**と定義する．

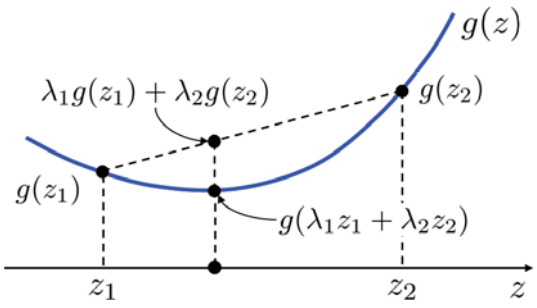

Fig. 2　Jensen の不等式

定理 1 (補助関数法)**.** 補助関数 $G(\theta,\bar{\theta})$ を，$\bar{\theta}$ に関して最小化するステップと，θ に関して最小化するステップ

$$\bar{\theta} \leftarrow \operatorname*{argmin}_{\bar{\theta}} G(\theta, \bar{\theta}) \tag{8}$$

$$\theta \leftarrow \operatorname*{argmin}_{\theta} G(\theta, \bar{\theta}) \tag{9}$$

を繰り返すと，目的関数 $D(\theta)$ の値は単調収束する．

ここで，距離関数 D が凸関数であれば次の Jensen の不等式を用いて補助関数を求めることができる．

定理 2 (Jensen の不等式)**.** 任意の凸関数 g，I 個の実数 $x_1,\ldots,x_I$，$\sum_i \lambda_i = 1$ を満たす I 個の正値の重み係数 $\lambda_1,\ldots,\lambda_I$ のもとで，

$$g\Big(\sum_i \lambda_i z_i\Big) \le \sum_i \lambda_i g(z_i) \tag{10}$$

が成り立つ *(図 2 参照)*.

例えば，スペクトログラム間距離を二乗誤差で測った場合，$\sum_k \lambda_{\omega,t,k} = 1$ を満たす正値の重み係数 $\lambda_{\omega,t,k}$ を用いて

$$\begin{aligned} J(\Theta) &= \sum_{\omega,t} \left(Y_{\omega,t} - \sum_k \lambda_{\omega,t,k} \frac{\alpha_k X_{k,\omega,t-\tau_k}}{\lambda_{\omega,t,k}} \right)^2 \\ &\le \sum_{\omega,t} \sum_k \lambda_{\omega,t,k} \left(Y_{\omega,t} - \frac{\alpha_k X_{k,\omega,t-\tau_k}}{\lambda_{\omega,t,k}} \right)^2 \\ &= \sum_k \sum_{\omega,t} \frac{1}{\lambda_{\omega,t,k}} \big(\lambda_{\omega,t,k} Y_{\omega,t} \\ &\qquad - \alpha_k X_{k,\omega,t-\tau_k} \big)^2 \end{aligned} \tag{11}$$

のような不等式が立てられ，右辺を $G(\Theta,\lambda)$ と置けば，$G(\Theta,\lambda)$ は目的関数 $J(\Theta)$ の補助関数としての要件を満たす．ここで，上述の不等式の等号

条件は $\lambda_{\omega,t,k} = \frac{\alpha_k X_{k,\omega,t-\tau_k}}{\sum_{k'} \alpha_{k'} X_{k',\omega,t-\tau_{k'}}}$ であるので，あとは補助関数法の原理に従い，

$$\lambda_{\omega,t,k} \leftarrow \frac{\alpha_k X_{k,\omega,t-\tau_k}}{\sum_{k'} \alpha_{k'} X_{k',\omega,t-\tau_{k'}}} \tag{12}$$

$$\alpha_k \leftarrow \underset{\alpha_k}{\operatorname{argmin}}\, G(\Theta, \lambda) \tag{13}$$

$$\tau_k \leftarrow \underset{\tau_k}{\operatorname{argmin}}\, G(\Theta, \lambda) \tag{14}$$

を繰り返していくことで $J(\Theta)$ を局所最小化する MIDI パラメータ Θ を得ることができる．式 (13) の更新式は，$\partial G/\partial \alpha_k = 0$ を解くことにより解析的に得られる．また，式 (14) では，τ_k の初期値の周辺の離散時刻の中から $G(\Theta, \lambda)$ を最小にする τ_k を探索することとする．先に挙げたほかの乖離度規準を用いた場合についてもパラメータ推定アルゴリズムが以上と同様に得られる．

2.2 時間軸方向での初期値の設定

今回は目的関数（式 1）の最適化に対して大域最適解を探索することが困難であるため，補助関数法という高速な局所最適解探索の手法を用いた．さらに楽譜情報から得られる機械的演奏を仮定した MIDI 信号を与え，正解に近いオンセット位置と音高を初期値として探索をすることで探索すべき解空間を大幅に狭めることができた．しかし，人間の演奏は常にテンポ変動を含むため楽譜上の音符列情報に対して時間方向の伸縮が加わった形で演奏される．ゆえに楽譜情報からの機械的演奏の MIDI 信号を与えても，実際の音響信号と MIDI 上の時間軸に対する同期をとらなければならないという課題が残っている．そこで，動的計画法によるマッチング（DP マッチング）の結果を実時間情報の初期値として利用した．これにより DP マッチングをせずに補助関数法による最適化を行った場合よりも，時間方向に関して推定誤りを起こしにくくなることが予想される．

しかし，スペクトログラム同士のマッチングを考えるとパワーの違いによる影響が大きく出てしまう．そこで，本報告では，Ellis の方法 [10] を利用した．この方法では，類似度を音楽音響信号のスペクトログラムの周波数成分のうち，MIDI 信号の各ノートの音高に対応する周波数成分の割合で表現するため，スペクトログラムを定数倍しても類似度が変化しないという特徴がある．

3 MIDI パラメータ推定実験

提案法で MIDI パラメータを推定できるかどうかを見るために，実演奏に対し推定を行った例を示す．

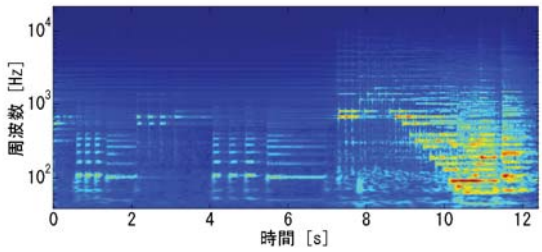

Fig. 3　音楽音響信号のスペクトログラム

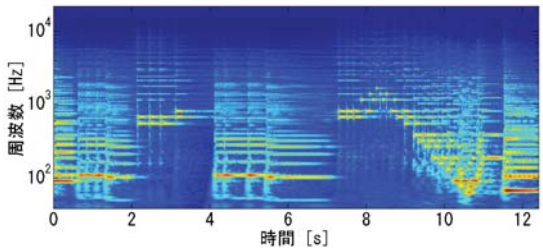

Fig. 4　DP マッチングにより推定された MIDI パラメータから生成される音響信号のスペクトログラム

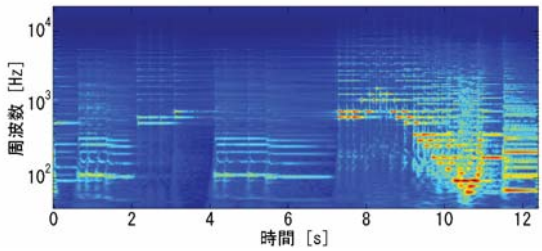

Fig. 5　補助関数法により推定された MIDI パラメータから生成される音響信号のスペクトログラム

実験の用いたデータは，RWC のクラシック音楽データベース [11] から，No. 28（L. v. ベートーヴェンのピアノソナタ第 23 番）の 12 小節目から 15 小節目を抜粋して利用した．スペクトログラムの計算は Gabor の Wavelet 変換を用い，そのパラメータは時間分解能が 8 ms，周波数分解能が 48 cent，最低中心周波数が 50 Hz とした．また，距離関数は I ダイバージェンスを用いて，補助関数法の反復回数を 30 回とした．MIDI パラメータは発音時刻（オンセット），消音時刻（オフセット），発音強度（ベロシティ）の 3 種類を考慮し，各反復ごとでオンセット・オフセット位置の最適値の検索範囲は，ともに反復前のオンセット・オフセット位置から ±400 ms とした．

図 3-5 はそれぞれ，フィッティングもとの音楽音響信号のスペクトログラム，DP マッチングで

推定したMIDIパラメータから生成される音響信号のスペクトログラム，DPマッチングで推定したMIDIパラメータを初期値として，補助関数法で推定したMIDIパラメータから生成される音響信号のスペクトログラムを示す．

図3と図5を比較すると音楽音響信号のスペクトログラムをよく近似できるMIDIパラメータが推定出来ていることが分かる．また，図3と図5を図4と比較すると，DPマッチングでは推定することが出来なかったベロシティが，補助関数法により推定出来ていることが分かる．

但し，図3と図4を比較すると，ノートが低音部のみに現れているときや，速いノートが連続する場合に，DPマッチングで推定誤りが生じていることが分かる．これは，Wavelet変換の低周波領域における時間分解能の低さや残響の影響と考えられる．更に，この結果を踏まえ，図5を見ると，DPマッチングにより生じた推定誤りが補助関数法による推定に強く影響を及ぼしていることが読み取れる．このため，今後の研究では，DPマッチングの精度の向上や，補助関数法を行う上でアニーリングなど初期値依存性を緩和させる手法の検討をする必要がある．

4　まとめ

本報告では，楽曲の楽譜とそれが実際に演奏された音楽音響信号が与えられている状況のもとで，音楽音響信号をMIDI信号で音響的に良く再現するよう，MIDIパラメータを自動チューニングする方法を提案した．時間領域では比較するのが困難であるこの問題に対し，MIDI信号から生成される音響信号と観測音響信号をスペクトログラム上の距離の最小化として捉え定式化した．さらに，膨大な解空間を探索しなければならないこの最適化に対して，元の楽譜から得られる機械的演奏を仮定したオンセット，音高情報とDPマッチングから得られる実時間情報を用いて解空間を効率的に探索する手法を提案した．実験ではベロシティ・オンセット・オフセットといったMIDIパラメータを推定することができ，本手法の有効性が示された．今後の研究では，MIDIパラメーターの種類を増やして実験を行うことを検討している．

謝辞　有益な助言を橘 秀幸氏（東大院･情報理工）からいただいた．本研究の一部は，文部科学省/学術振興会科学研究補助費 課題番号 (23240021) から補助を受けて行われた．

参考文献

[1] T. H. Kim, S. Fukayama, T. Nishimoto and S. Sagayama. Polyhymnia: An automatic piano performance system with statistical modeling of polyphonic expression and musical symbol interpretation. In Proc. NIME, pp. 96-99, 2011.

[2] C. Raphael, “A hybrid graphical model for aligning polyphonic audio with musical scores,” in Proc. ISMIR, pp. 387–394, 2004.

[3] P. Peeling, et al., “A probabilistic framework for matching music representations,” in Proc. ISMIR, pp. 267–272, 2007.

[4] 松本, 西本, 小野, 嵯峨山, “楽譜からの音楽音響信号生成モデルに基づく楽譜と音響信号の詳細な整合,” 音講論 (秋), pp. 847–848, 2007.

[5] A. Maezawa, et al., “Polyphonic audio-to-score alignment based on bayesian latent harmonic allocation hidden Markov model,” in Proc. ICASSP, pp. 185–188, 2011.

[6] 武内, 亀岡, 齋藤, 深山, 嵯峨山, “動的時間伸縮調波時間構造化クラスタリング (TW-HTC) による音楽音響信号と楽譜の自動整合,” 日本音響学会研究資料, pp. 19–24, 2012.

[7] 亀岡, 鎌本, 原田, 守谷, “予測誤差の golombrice 符号量を最小化する線形予測分析,” 信学論, vol. J91-A, no. 11, pp. 1017–1025, 2008.

[8] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. in Adv. NIPS, pp. 556–562, 2000.

[9] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama. Complex NMF: A new sparse representation for acoustic signals. in Proc. ICASSP, pp. 3437–3440, 2009.

[10] D. P. W. Ellis. Aligning midi scores to music audio, web resource. `http://www.ee.columbia.edu/~dpwe/resources/matlab/alignmidiwav/`.

[11] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical, and jazz music database. in Proc. ISMIR, pp. 287–288, 2002.