5C-02

時間周波数分解能の異なる2つのスペクトログラムに対する並列NMFを用いた演奏詳細解析*

保利武志　中村和幸　嵯峨山茂樹
明治大学先端数理科学研究科

1　はじめに

本論文では，音楽演奏の自動表情付けを実演奏データから学習するための，微細なオンセット時刻の揺らぎを詳細に解析する手法について述べる．

演奏表情は緩急や強弱、スラーやスタッカート等のアーティキュレーションとして演奏に反映されるが，実演奏データから演奏表情学習に必要な微細なオンセット時刻等を抽出するためには，高精度にその変化を捉える必要がある．CD音源のスペクトルから音高を推定する手法はこれまで様々な提案が為されているが，特にNon-negative matrix factorization (NMF) をベースとした手法は，楽器音特有の調波構造モデルを基底に組み込みやすく，また振幅スペクトルの加法性仮定や波形の加法性との相性と相まって高精度な推定を実現している．しかし一般に，短時間フーリエ変換 (STFT) において，時間分解能と周波数分解能は不確定性原理に基づく解析フレーム長のトレードオフな関係があることから，各音高に対応する詳細なオンセット時刻を推定するための時間分解能に対し，十分な周波数分解能を担保することは難しい．

本研究ではピアノ演奏を対象として，そのトレードオフな関係を解消して高精度な音高及びアクティベーションの推定を実現するために，高時間 / 高周波数分解能な2つのスペクトログラムを用いて互いの基底情報とアクティベーションを参照し合う並列NMF(CNMF)[1]を用いた微細なオンセット時刻を推定する方法を提案する。

2　CNMFを用いた演奏の詳細解析

2.1　楽譜演奏と表情付き演奏

本論文では楽譜通りに弾かれる演奏を” 楽譜演奏”，演奏表情が付与された演奏を” 表情付き演奏” と呼称する．本研究の目的は演奏表情学習のための特徴量 (本論文では特にオンセット時刻) 抽出であり，楽譜情報は既知である．図1に示すように，楽譜演奏は基本的に十六部音符単位等でクオンタイズされる結果垂直方向のオンセット時刻が揃う傾向にあるのに対し，演奏者の個人性が顕著に現れる表情付き演奏では演奏者独特の癖や解釈に応じてオンセット時刻が均一に揃わないことが多い．従って，全体的なタイムアライメントを取った上で各音高にどの程度の揺れがあるのかを検出する必要がある．

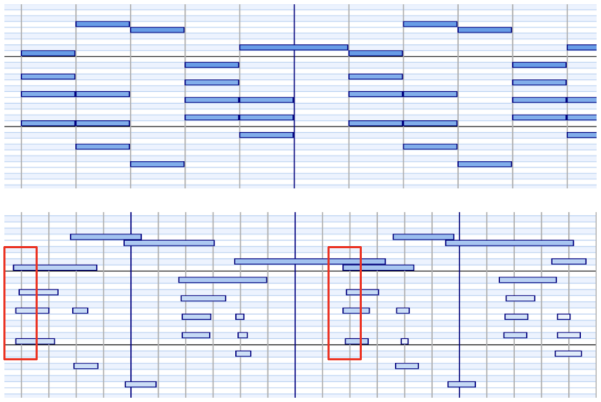

図 1. Chopin, Ballade No.4, Op52 における楽譜演奏 (上図) と表情付き演奏 (下図) のピアノロールの一部．表情付き演奏では垂直方向のオンセット時刻に微細な揺れが生じる

2.2　提案手法

CNMFで得られた各基底がそれぞれの音高と対応していれば，アクティベーションの推移を解析することでオンセット時刻を得られる．その際，オンセット時刻の微細な揺れを検出する必要があることから，時間解像度は大きな値を取る必要 (例えばテンポ120bpmである場合の十六分音符のdurationは125msであり，詳細解析のためにはこれ以上の短いフレーム長による分析が必要) がある一方で，周波数解像度は最低限各音高の周波数に分解できる解像度 (例えば$C2$：65.4Hzと$C2\#$：69.3の場合は約4Hzの周波数分解能) を保証する必要がある．

*Concurrent nonnegative matrix factorization using multi-resolution spectrograms for further analysis of music signals
Takeshi HORI, Kazuyuki NAKAMURA and Shigeki SAGAYAMA
Advanced Mathematial Sciences, Meiji University

2.3 並列 NMF: Concurrent NMF, CNMF

スペクトログラム $Y^{\in\omega\times t}$ を低ランクな基底行列 $H^{\in\omega\times k}$ とアクティベーション $U^{\in k\times t}$ の積で近似することを考える．Y と HU の距離尺度に I ダイバージェンスを用いた場合，

$$\begin{aligned} &\mathit{minimize}\quad \mathcal{I}(\theta) \\ &=\sum_{\omega,t}\left[Y_{\omega,t}\log\frac{Y_{\omega,t}}{\sum_k H_{\omega,k}U_{k,t}}-\left(Y_{\omega,t}-\sum_k H_{\omega,k}U_{k,t}\right)\right], \\ &\mathit{subject\ to}\quad \sum_\omega H_{\omega,k}=1,\quad H_{\omega,k}\geq 0,\quad U_{k,t}\geq 0,\quad \theta=\{H,U\} \end{aligned} \tag{1}$$

のように尺度距離最小化問題として定式化できる．基底ベクトルはスケールの任意性を回避するために正規化される．式 1 は Jensen の不等式を用いた Majorize-Minimization(MM) アルゴリズムにより，H,U は乗法更新によって推定可能である．CNMF ではさらに，異なる解析フレーム長比較に起因する共通の周波数ビン及び各フレームのアクティベーションに関する類似性 $(\mathcal{R}_H(\theta),\mathcal{R}_U(\theta))$ と，アクティベーション L_p ノルム正則化 $(\mathcal{S}(\theta))$ を用いて，

$$\begin{aligned} \mathit{minimize}\quad & \mathcal{J}(\theta)=\sum_n \mathcal{I}^n(\theta)+\mu_H\mathcal{R}_H(\theta) \\ & \qquad +\mu_U\mathcal{R}_U(\theta)+\lambda\mathcal{S}(\theta), \\ \mathit{subject\ to}\quad & \sum_\omega H_{\omega,k}=1,\ H_{\omega,k}\geq 0,\ U_{k,t}\geq 0, \\ & \theta=\{H,U\},\ n=\{S,L\},\ \mu_H,\mu_U,\lambda\geq 0 \end{aligned} \tag{2}$$

として定式化され，これは同様に MM アルゴリズムで最適化できる．S,L はそれぞれ短 / 長フレーム解析を表し，μ_*,λ は各正則化項に対する重みである．また基底に関しては，調波成分を考慮した打ち切り正規分布の混合による近似を初期値として用いることで，単音別のスペクトルが得られる [2]．上式から得られる更新式は，高周波数分解能な基底情報を用いて高時間分解能なアクティベーションを更新するような，互いに参照し合う形となる．

3 評価実験

CNMF を用いた詳細オンセット時刻の推定実験を行った．2 つのスペクトログラムにおける解析フレーム長はそれぞれ $64ms, 256ms$，フレームシフトはハーフオーバーラップとして STFT し，各パラメータはそれぞれ $\mu_H=0.5, \mu_U=2.0, \lambda=1.0, p=0.5$ とした．

図 2 は International Piano-e-Competition [3] における Chopin の Ballade No.4, Op.52 の表情付き演奏の一部を解析した結果である．基底数は楽譜情報を参考に 21＋1(音立ち上がりの広域周波数を吸収するための基底) とし，各基底ベクトルは調波構造 $h(f)$ として，n 倍音成分 f_n が基本周波数 f_0 に対して，

$$\frac{h(f_n)}{h(f_0)}=(n+1)^{-1.5}, \tag{3}$$

となるよう初期値を定めた．図 2 の左図は約 15 秒程度に対する CNMF の解析結果 (に楽譜上現れなかった音高を追加したもの) であり，右図は枠部分を抽出し拡大したものである．左図を見るとそれぞれのアクティベーションは各音高に対応するよう推定できており，また右図を見ると CNMF によって微細なオンセット時刻の揺れが検出できていることがわかる．

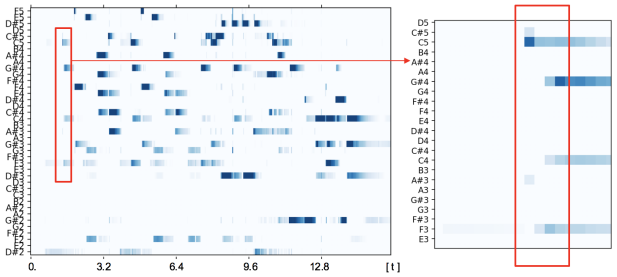

図 2. Chopin, Ballade No.4, Op.52 に対する CNMF を用いた詳細解析結果．左図は約 15 秒程度の解析結果で右図はその一部を抽出拡大したもの

4 おわりに

本研究では演奏表情学習のための詳細なオンセット時刻推定について，時間分解能と周波数分解能の解析フレーム長に関するトレードオフを解消をした CNMF による詳細解析とその評価を行った．楽譜に基づく音高情報と実験結果から，各基底がそれぞれの音高に対応するよう推定され，またその基底に基づく詳細なアクティベーション (オンセット時刻) の検出が可能であることが示された．

今後は CNMF の解析結果に基づき楽譜演奏とのタイムアライメントを取り解析箇所を限定することで，より細かい解析フレーム長による詳細なオンセット時刻を抽出したい．また，今回焦点としたオンセット時刻以外にも重要な duration や強弱などのモデルもアクティベーションの形状として含めることでより多くの特徴量を抽出し，ニューラルネットベースで学習することで実際に演奏表情付与を行いたい．

参考文献

[1] 落合和樹他, ”時間周波数分解能の異なるスペクトログラムの並列 NMF による多重音解析,” 研究報告音楽情報科学 (MUS), Vol.2011, No.5, pp.1-6, 2011.

[2] Kameoka, et al.,“A multipitch analyzer based on harmonic temporal structured clustering,” IEEE Trans. on Audio, Speech, and Language Processing, vol.15, pp.982-994, 2007.

[3] YAMAHA, et al., “International Piano-e-Competition,” http://www.piano-e-competition.com (2018.1).