The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

RBMを用いた楽器音基底と演奏情報への分離による多重音解析

Polyphonic music factorization into sound basis and activation using RBM

荒川賢也　　中鹿亘
Kenya Arakawa　　Tôru Nakashika

*1電気通信大学
The University of Electro-Communications

Recently, music studies based on deep learning that require a large amount of input have been garnering attention increasing. Along with that, the task of generating accurate scores from audio data is also important. Although NMF is often used for music factorization into sound basis and activation, there is room for improvement and many methods have currently being proposed. In this paper, we propose method of polyphonic music factorization using RBM. RBM is stochastic model and outputs binary-valued latent features, which is suitable for music score notation. Furthermore, we also propose sparse-RBM in order to settle cross cancel problem. In conclusion, our proposed method showed better accuracy than NMF.

1. はじめに

近年，音楽研究において自動作曲など発展的な研究で膨大な楽譜データが必要とされる．そのため波形信号から楽譜を自動生成するタスクもデータ確保のために重要な問題であると考えられる．しかし楽器音は基本周波数の他に倍音を含んでいるため，和音を含んだ音楽の解析は非常に難しい問題でもある．音声データから事前知識を用いずに楽器音基底行列 (音階) とその演奏情報行列 (楽譜) に分離する楽器音分離においては non-negative matrix factorization(NMF) がよく用いられている [1]．しかし分離精度には未だに向上の余地があり，その他にも多くの楽器音分離アルゴリズムが提案，議論されている．

生成モデルの restricted Boltzmann machine(RBM) は NMF と同様の行列分解を行うことができる．二つのアルゴリズムの大きな違いとして NMF が決定的に解を求めるのに対して，RBM は確率的に解を求めるという点，NMF は演奏情報行列が連続値で出力されるのに対して RBM はバイナリ値で出力される点が挙げられる．音楽などの感性に関わる曖昧なものに対しては確率モデルの方が適していると考えられる．例えば機械的に同期した演奏ではなく生演奏から楽譜を生成する場合，決定的に解を導く NMF よりも確率分布による RBM の方が入力のわずかな差異に対しても柔軟な結果が期待できる．また，楽譜の表現としては連続値よりもバイナリ値の方が適していると考えられる．本研究ではそのようなアルゴリズムの違いによる入力への柔軟性及び出力形式の違いから RBM を用いて音楽の波形信号を楽器音基底行列とその演奏情報行列に分離する実験を行い，その性能を NMF と比較，検討する．

2. 従来手法:NMF による多重音解析

非負の行列 $\mathbf{X} \in R^{\geq 0, M\times N}$ を式 (1) のコスト関数を最小化することで $\mathbf{W} \in R^{\geq 0, M\times R}$ と $\mathbf{H} \in R^{\geq 0, R\times N}$ の積の形への近似を行う．

$$C = \|\mathbf{X} - \mathbf{W}\cdot\mathbf{H}\|_F \tag{1}$$

ここで $\|\cdot\|_F$ はフロベニウスノルムを示す．

音楽の楽器音分離に適用する場合，入力 $\mathbf{X}$ に振幅スペクトル (M:スペクトルビン数,N:時間ビン数)，R にランク数 (分離する音数) を与えることで $\mathbf{W}$ が楽器音基底行列，$\mathbf{H}$ がその演奏情報行列として分離される [1]．

3. 提案手法:RBM による多重音解析

RBM は可視層と隠れ層からなる二層のネットワークであり，パラメータとして可視素子 $\mathbf{v}$，隠れ素子 $\mathbf{h}$ のそれぞれのバイアス $\mathbf{b}$, $\mathbf{c}$，各層間の素子の結合の重み $\mathbf{W}$ を持っている．

3.1 Gaussian-Bernoulli RBM

RBM の結合確率は式 (3) の正規化定数 Z を用いて次のように表せる．

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp(-E(\mathbf{v},\mathbf{h})) \tag{2}$$

$$Z = \sum_{\mathbf{v},\mathbf{h}} \exp(-E(\mathbf{v},\mathbf{h})) \tag{3}$$

式 (2) に示す結合確率について最尤推定を行うことでパラメータを調整し，入力を再現するような確率分布を求める．本研究では入力に振幅スペクトルを用いるため可視素子は連続値，隠れ素子はバイナリ値を扱うことができる Improved Gaussian-Bernoulli RBM(IGBRBM)[3] を用いる．

IGBRBM のエネルギー関数 E は，

$$E(\mathbf{v},\mathbf{h}) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_j c_j h_j - \sum_{i,j}\frac{v_i}{\sigma_i^2} W_{ij} h_j \tag{4}$$

で定義される．ここで σ_i は可視素子 v_i の分散のパラメータを示す．以降，本稿では IGBRBM を RBM として表記する．

RBM を楽器音分離に適用する場合は可視素子数がスペクトルビン数，隠れ素子数が基底数のネットワークに対して入力として振幅スペクトルを与えることで重み $\mathbf{W}$ が楽器音基底行列として学習され，隠れ層 $\mathbf{h}$ に演奏情報が出力される．

3.2 sparse-RBM

RBM を用いて楽器音分離を行う場合，楽器音基底が負の値を取ることによってクロスキャンセルが起こり，特徴的周波数が打ち消しあう可能性がある．これを回避するために sparse-RBM を用いる [4]．sparse-RBM はコスト関数として対数尤

度 J に制限項を足した次式 (5) を用いることで隠れ層 $\mathbf{h}$ をスパースな状態に制限し，演奏情報行列が同時に 1 になる回数をなるべく少なくする．

$$C = -J + \lambda \sum_{\mathbf{h}} |p - \mathbb{E}[\mathbf{h}|\mathbf{v}]|^2 \tag{5}$$

ここで $\mathbb{E}[\cdot]$ は期待値を示し，p と λ はそれぞれスパース制限の強さを決定するハイパーパラメータである．

このコスト関数を最小化するパラメータを学習する際は第一項の対数尤度と第二項の制限項は分けて計算を行う．つまり通常の RBM の学習を行ったのちに第二項に関してパラメータを更新するようにする．また，第二項に関するパラメータの更新では隠れ素子のバイアス c のみを更新する．

4. 実験

4.1 実験条件

本研究では RWC 研究用音楽データベース (クラシック音楽) を利用した．MIDI データをピアノ音源を用いて wav データに書き出し，それをフーリエ変換することで振幅スペクトルを得る．これを平均 0，分散 1 に正規化したものを RBM の入力として用いた．正解ピアノロールのノート数を C，演奏情報行列のノート数を A，正解ピアノロールと演奏情報行列で異なる音がアクティベーションしているノート数を D とした時，以下の式 (6) で表される正解率を用いて評価を行った．

$$Accuracy = \frac{C + A - D}{C + A} * 100 \tag{6}$$

はじめに suparse-RBM のハイパーパラメータに関してランク数 8 で 8 s の短く単純な音源とランク数 21 で 32 s の長く複雑な音源の二つの音源を用いて実験を行った．先行研究 [4] に習い $\lambda = 1/p$ とした．p の値を変更しながら一つのパラメータに対して 10 回楽器音分離を実行，評価してその平均を出力した．

次に RWC 研究用音楽データベースより 9 つの音源を用いて実データを用いた従来手法との比較実験を行った．それぞれの音源に関して NMF と sparse-RBM を用いて各 10 回楽器音分離を実行，評価してその平均を出力した．sparse-RBM のハイパーパラメータは音源ごとに調節した．

4.2 ハイパーパラメータに関する実験

実験結果を図 1 に示す．最大となったのはランク 8 の音源では $p = 0.06$，ランク 21 の音源では $p = 0.42$ の時であった．ランク 8 の音源の方がスパースであることから p の値が小さいほどより強く演奏情報行列のスパース性を保証する事が確認できる．また図 1 よりハイパーパラメータは分離精度に大きく影響するため音源の複雑さに対して適切に設定される必要があることがわかる．

4.3 従来手法との比較

実験結果を表 1 に示す．提案手法は NMF に対し平均の分離精度で上回った．特にランク数の低い単純な音源 (RM-C27) において NMF よりもかなり高い分離精度を示した．これはスパース制約によってクロスキャンセルが是正され，楽器音基底が高い精度で分離されたためであると考えられる．音基底で演奏される頻度に偏りがあるものは全体的に分離精度が低くなっている．これは情報量が少ない楽器音基底をうまく分離できなかった場合に他の楽器音基底に与える影響が大きいためであると考えられる．ランクの高い複雑な音源では提案手法の

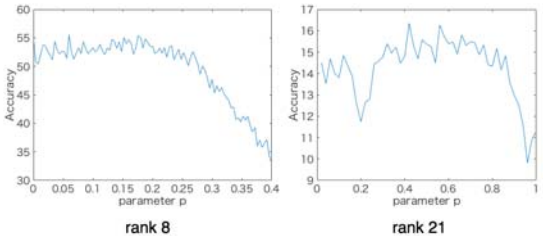

図 1: sparse-RBM のパラメータ p の変化による分離性能の推移

分離精度は NMF と同等かあるいは劣る結果となったのは複雑な音源ではスパース制約が弱くなることでクロスキャンセルが発生してしまうためであると考えられる．これらの結果より sparse-RBM の分離精度は音源の複雑さ及びデータの偏りに依存していると考えられる．演奏情報行列のスパース制限以外の方法でクロスキャンセルを抑えることで複雑な音源における RBM の分離精度を改善することができると考えられる．

表 1: 従来手法との比較実験結果

filename	rank	time(s)	NMF(%)	sparse-RBM(%)
RM-C27	9	33	32.4	59.3
RM-C31	18	24	15.7	13.1
RM-C30	20	32	13.1	10.8
RM-C23A	21	30	18.7	17.4
RM-C23B	25	48	14.7	15.3
RM-C26	27	28	27.2	23.8
RM-C35A	32	31	13.3	11.4
RM-C23C	35	55	26.6	24.1
RM-C29	41	32	12.3	12.1
average	25.3	34.8	19.3	**20.8**

5. まとめ

本稿では RBM を用いて波形信号から楽器音基底行列と演奏情報行列へ分離する方法を提案した．提案法は NMF による同様の楽器音分離と比較して平均してわずかに良い結果を示した．また RBM を楽器音分離に適用するメリットとして演奏情報行列の出力が NMF は連続値であるのに対して RBM はバイナリ値で表現される点が挙げられる．

今後の展望としては RBM における W の非負性を sparse-RBM と異なる拡張方法で実現すること，同一音階の複数楽器を含む音源及び，同一楽譜における複数の生演奏音源での分離の実験などがある．

参考文献

[1] P. Smaragdis and J.C. Brown:“Non-negative matrix factorization for polyphonic music transcription” Applications of Signal Processing to Audio and Acoustics, IEEE Workshop. (2003)

[2] Geoffrey E. Hinton: “A Practical Guide to Training Restricted Boltzmann Machines”,Lecture Notes in Computer Science, vol. 7700. (2012)

[3] KyungHyun Cho *et al*:“Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines”,Lecture Notes in Computer Science, vol. 6791. (2011)

[4] Honglak Lee *et al.*: “Sparse deep belief net model for visual area V2”, Neural Information Processing Systems 20. (2007)