3R-10

単音の音量ダイナミクスを共有化したNMFによる楽器パート分離

田島 照久 † 阪上 大地 ‡ 糸山 克寿 ‡ 奥乃 博 ‡

† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

近年, 歌声合成ソフト「初音ミク」などの Desktop Music (DTM) ソフトウェアの充実により, 誰でも手軽に音楽制作ができるようになり、 さらに, 製作したコンテンツをニコニコ動画などの Consumer Generated Media (CGM) サイトへ投稿することで活発な交流ができるようになった. CGM では既存の音楽素片を複数組み合わせて編曲を行う場合もあり, パート音を素材としたいと考える人は多い. このような音の素材化には合奏音から楽器パートごとに精度良く分離することが必要となる.

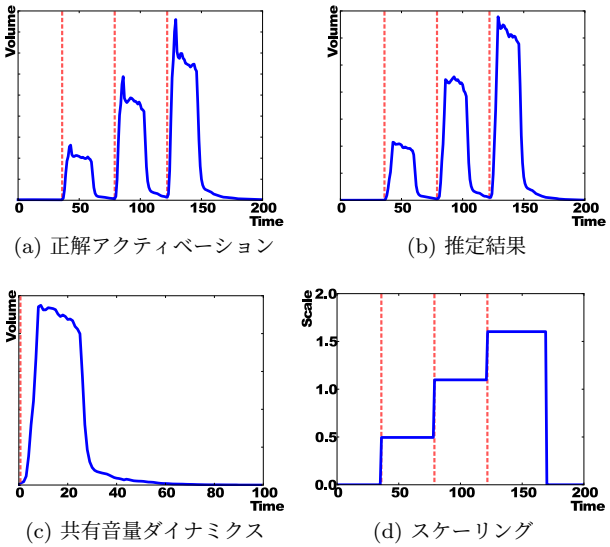
本稿で提案する楽器パート分離では非負値行列因子分解 (Non-negative Matrix Factoraization; NMF) を使用し, 合奏音のスペクトログラムを楽器パートごとのスペクトログラムに分離する. NMF は音響信号に応用すると, 観測されたスペクトログラム Y を時不変なスペクトル構造である基底 H と, 各時刻での基底の重みを表すアクティベーション U の積に分解する手法である. ただし、NMF はスペクトル構造が異なる音をすべてばらばらに分解してしまうので, そのままでは楽器パートごとの分離はできない.

NMF は観測データから構成因子を推定する逆問題であるので, 適切な解を得るには目的に合った制約条件を仮定する必要がある. 楽器パート分離では, 楽器ごとのスペクトログラムを得ることが目標である. 制約条件として, Ewert ら [1] は楽譜に基づくスペクトル・時間マスクを使用している. Harmonic NMF [2] は基底に楽器由来の調波構造を初期値としている. Convolutive NMF [3] は基底に数フレーム分のスペクトルを用いることで周波数構造の時間変動を許容している. ただし、いずれの方法も単音の音量ダイナミクスが各単音間で類似する性質を用いていない. 単音間類似性とは, 弾く瞬間の強さによって音が決まる楽器では, 楽曲の序盤に鳴る単音も, 終盤に鳴る単音も類似した音量ダイナミクスを持ち, それぞれの単音の発音開始直後の音量比と発音終了直前の音量比が同じであることを意味する. また, 各パートを精度よく分離するためには, 単音間類似性は多くの楽器で成立するように設計しなければならない.

本稿では, 音響信号と対応付けられた MIDI データから作成した演奏される楽器と音高の一覧に対し, 楽器・音高の各ペアが単一の時不変なスペクトルに従うと仮定する. これを基底と呼ぶ. また, 各基底の単音の音量変化についてもある種のテンプレートに従うと仮定する. このテンプレートを共有化された音量ダイナミクスと呼ぶ. 各単音の音量のばらつきを許容するため, スケーリングパラメーターを導入する. まとめると, 本稿でモデル化する楽器音は, 次の条件を満たすことを想定する.

1. 音量ダイナミクスが音の発音時からの経過時間のみに依存し, 発音時刻に依存しない.
2. 同一音高であれば任意の強さの単音のダイナミクスは 1 つのダイナミクスのスケーリングで表される.

例を用いて説明する. 図 1 (a) のようなアクティベーションは, 図 1 (c) の発音開始時刻を起点とする共有音量ダイナミクスと, 図 1 (d) の相対音量との積を発音区間に適用することで推定し, 図 1 (b) の分離結果を得る.



(a) 正解アクティベーション (b) 推定結果

(c) 共有音量ダイナミクス (d) スケーリング

図 1: モデルの例. 青線は音量変化, 赤線は発音タイミングを示す. (a) のような音量変化を含んだパートを推定するには, 共有音量ダイナミクス (c) を, スケーリング (d) により補正し, 推定結果 (b) を得る.

以下ではこの単音間類似性を用いて, 楽曲中の複数の単音のダイナミクスを共有化し, アクティベーションに制約を加えることで分離精度の向上をめざす.

2. 音量ダイナミクスの共有化

時刻 t, 周波数 f, 基底 k のスペクトルを s^k_{tf} とする. 基底を h^k_f, アクティベーションを u^k_t で表すと, $s^k_{tf} = h^k_f u^k_t$ となる. 本手法では共有音量ダイナミクスを用いて u^k_t をモデル化する.

2.1 共有化モデル

基底 k の単音の共有音量ダイナミクスを u^k_l, r 番目の単音のスケーリングを v^k_r とする. ここで, l は各単音の発音後の経過時間である. さらに, $S^k_t : t \mapsto l$ は楽曲の開始時点から時刻 t の位置でちょうど鳴っている単音について, その単音が発音されてからの時刻 l を表す. また, スケーリングは各単音ごとに設定したいので, $R^k_t : t \mapsto r$ によって楽曲の開始時点から時刻 t の位置で鳴っている単音ををを r 番目の単音であると表す. これによりスペクトルは $s^k_{tf} = h^k_f u^k_{S^k_t} v^k_{R^k_t}$ となる.

2.2 更新式の導出

NMF の更新式導出は Bayesian NMF [4] をもとに行う. 本稿で提案するモデルを用いた u についての更新式を導出する. スペクトルは Poisson 分布とし, 各変数の事前分布は共役事前分布となるよう Gamma 分布とする.

$$p(s^k_{tf}|H,U,V) = \mathcal{PO}(s^k_{tf}|h^k_f u^k_{S^k_t} v^k_{R^k_t}) \quad (1)$$

$$p(u^k_l) = \mathcal{G}(a_0, b_0) \quad (2)$$

表 1: 制約別 SNR (dB)

dataset	SD-HN	SD	HN	baseline
Original	**−1.19**	−2.32	−3.50	−4.51
Random	**−1.39**	−2.56	−3.95	−5.17

U の変分事後分布は, $n_t^k \equiv \sum_f \mathbb{E}[s_{tf}^k], H_k \equiv \sum_f \mathbb{E}[h_f^k]$ を定義し, 更に $l = S_t^k$ となる t の集合を T^k を用いて次式になる.

$$\ln q^*(u_l^k) = \sum_{T^k} n_t^k \ln u_l^k - H_k \sum_{T^k} \mathbb{E}[v_{R_t^k}^k] u_l^k + \ln p(u_l^k)$$

式 (2) の事前分布を用いると更新式は次式になる.

$$a_l^k = a_0 + \sum_{T^k} n_t^k \ , \quad b_l^k = b_0 + H_k \sum_{T^k} \mathbb{E}[v_{R_t^k}^k] \qquad (3)$$

3. 実験及び考察

本手法の有効性を調べるため, 評価実験を行った. なお, 音楽音響信号は MIDI データから生成した.

3.1 比較対象

4 つの異なる制約を用いた NMF の性能を比較する. EM アルゴリズムの反復回数はそれぞれ 100 回とした.

1. baseline : 楽器パート分離のベースラインとして, 休符部分は音量が小さいとして, アクティベーションにバイナリ (0/1) マスクをかけた.
2. HN (HarmoNic) : ドラム以外の基底に Ewert らの手法 [1] を用いて 1 つの調波構造を与え, アクティベーションはノートオフ後に音量が指数減衰する制約をかけた.
3. SD (Shared Dynamics) : 本稿で提案した, アクティベーションの共有化制約を加えた.
4. SD-HN : 上記 SD と HN を組み合わせたもので, ドラム以外の基底に調波構造を与え, アクティベーションの共有化制約を加えた.

3.2 実験条件

実験には RWC 音楽データベース [5] からポピュラー楽曲 No.1–10 の 10 曲, 各曲先頭 30 秒を使用した. 各楽曲は MIDI 音源 Roland SD-90 を用いて, 44.1kHz, 16bit, モノラルで録音した. STFT により時間周波数表現に変換した. 楽曲の構成楽器による性能差を調べるため, 以下の 2 つのデータセットを用意した.

1. Original : 提供されている MIDI データのままの楽器を用いた. (10 曲 ×1 パターン)
2. Random : 各曲で使用されているドラム以外の Program No. を重複なくランダムに書き換え, 構成楽器を変化させた MIDI を用いた. (10 曲 ×10 パターン)

3.3 実験方法

前述の各楽曲に対し, 比較対象の 4 つの NMF でパート分離を行った. 性能の指標として Signal-to-Noise Raito (SNR) を用いた. 分離を行なって得られた楽器別音高別のスペクトログラムを楽器別に合計したものと, パートごとに録音した正解データのスペクトログラムとの比較で周波数領域においてパートごとの SNR を計算した. 平均 SNR はそのパートごとの SNR を各パートの音量で重み付けした加重平均とした. 加重平均を用いる理由は, 小音量のパートは分離の重要度が小さいわりに, 少量のノイズの影響を受けやすく, 単純平均では手法の性能を正しく評価できないためである.

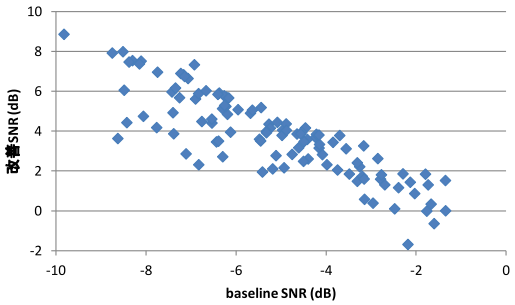


図 2: SNR の改善傾向

3.4 結果と考察

データセット別, 制約別に 10 曲の SNR を平均した結果を表 1 に示す. 実験の結果, 提案した SD-HN, つまり単音の音量ダイナミクスを共有化し, 基底に調波構造マスクを与えた場合に性能が向上することを確認できた. また, HN よりも SD が高性能であるため, 基底に対する調波マスクより, 音量ダイナミクスの共有化のほうが性能向上への寄与率が大きいことがわかった.

提案法が有効な入力データの特徴を調べるため, 横軸に baseline の SNR, 縦軸に baseline と SD-HN の SNR 差をとったグラフが図 2 である. 図中の点は全 10 曲 ×11 パターンの各曲各パターンに対応する. 図より baseline での SNR が悪い曲ほど提案法によって SNR が改善しているように見える. 相関係数を求めると −0.859 と強い相関があることが示される. つまり, 従来法での分離が難しい曲に対して提案法の効果が高いと示された. 構成楽器にかかわらずこの傾向が見られることから, 提案法が特定の楽器に特化している手法ではなく, また, パート分離は分離対象楽器以外の楽器をうまくモデル化することでも性能が向上するので, 本手法は多くの楽器に適用できると考えられる.

4. おわりに

本稿では, 音楽音響信号とそれに対応する楽譜情報から, NMF のアクティベーションに単音の音量ダイナミクスを共有化する制約を加え, 分離精度の向上を確認した. 今後の改良としては, 提案法では単音中の音量変化が単音間で類似するという仮定を基にしているが, 例えばトランペットなどの単音中も自由に音量を変えられる楽器音をモデル化し, 組み込むなどが考えられる. なお, 本研究の一部は科研費 (S) No.24220006, 若手 (B) No.24700168 の支援を受けた.

参考文献

[1] S. Ewert and M. Müller: Using score-informed constraints for NMF-based source separation, *ICASSP 2012*, pp. 129–132, March 2012.

[2] S.A. Raczyński, N. Ono, S. Sagayama: Multipitch analysis with harmonic nonnegative matrix approximation, *ISMIR 2007*, pp. 381–386, Sep. 2007.

[3] P. Smaragdis: Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs, *ICA 2004*, pp. 494–499, Sep. 2004.

[4] A.T. Cemgil: Bayesian inference in non-negative matrix factorisation models, Technical Report CUED/F-INFENG/TR.609, University of Cambridge, July 2008.

[5] 後藤 真孝, 橋口 博樹, 西村 拓一, 岡 隆一: RWC 研究用音楽データベース, 情報処理学会論文誌, Vol.45, No.3, pp. 728–738, March 2004.