同一楽曲に対する多数の歌唱の音高推移分布の可視化

近藤 芽衣　(指導教員：伊藤 貴之)

1　はじめに

同一楽曲に対して多数の歌唱者がソーシャルメディアなどに自分の歌唱作品を公開する機会が近年増えている．このような歌唱群の癖や個性を理解するための一手段として，同一楽曲に対する歌唱者群の歌い方を可視化する手法を提案する．この手法では，同一楽曲に対する多数の歌唱群の音響データからそれぞれの音高（基本周波数：以下 F0 と称する）の推移を抽出し，その分布を可視化する．また視覚表現の手段として，時刻および周波数の対数値を 2 軸とする 2 次元ヒストグラム画像を生成し，これに適応 2 値化・ラベリングといった画像処理手法を適用している．本研究ではこの可視化手法と可視化結果を紹介したのちに，一般的な歌唱演奏音源から無伴奏歌唱音響データを生成する手順，およびその可視化結果を示す

2　時系列データの可視化

歌唱の音高の推移は時系列データとして扱うことが可能であり，汎用的な時系列データ可視化手法を適用することが可能である．

ここで n 個の標本がそれぞれ m 個の時刻における実数値を有する時系列データがあるとする．このようなデータに関する多くの可視化手法は以下のいずれかのアプローチを有する．なお，以下での「実数値」は F0 値に,「密度」は近い音高を有する歌唱者の人数に対応する．

1. 一方の座標軸に m 個の時刻，他方の座標軸に実数値を割り当てた折れ線グラフや散布図.
2. 1. の折れ線や点群を密度関数に置き換えて，密度を各画素の明度や色相に変換したヒートマップで表現したもの.
3. 一方の座標軸に m 個の時刻，他方の座標軸に n 個の標本を割り当てたマトリクスに対して，実数値を各画素の明度や色相に変換したヒートマップで表現したもの.

これらのアプローチの各々にはいくつかの問題点がある．1. に示した折れ線グラフや散布図には，画面上の描画物の過密状態が引き起こす Visual Cluttering と呼ばれる視認性の低下が避けられない．また可視化結果からのデータ読み取りにおいて色の識別能力は高くないことが知られており，3. に示したヒートマップでは実数値を正確に読み取れない可能性がある．以上により本研究では 2. に示す「密度関数のヒートマップ」というアプローチをとることにする．

3　提案手法

本章では提案手法の各処理を手順に沿って示す.

3.1　一般的な楽曲音源からのデータ生成

伴奏とあわせて録音された一般的な歌唱音源から無伴奏データを生成することで，幅広い楽曲に対して可視化が可能になる．以下の手法を用いて伴奏付き歌唱音源から無伴奏歌唱音源を抽出する手法を開発した．この処理によって生成された無伴奏歌唱音源の F0 を推定することで，幅広い楽曲への適用が可能となる．

3.1.1　時刻及びキーオフセット推定

伴奏付き歌唱音源と伴奏音源を Constant–Q 変換によりスペクトログラムに変換する．この 2 音源の時間と周波数の二次元配列の相互相関の最大値を求めることにより，音源の始まる時刻やキーのずれを検出する．この時，相互相関を求める範囲には歌唱が含まれていないことが望ましい．これによりサビ始まりの曲以外は音源の始まりから 5 秒程度を用いる．

3.1.2　音量オフセット推定

時刻のずれを修正した伴奏付き歌唱音源と伴奏音源から STFT によって振幅スペクトルに変換したものをそれぞれ $S1$，$S2$ とする．この二つの音源の音量差を考慮して減算するため，無伴奏歌唱音源の振幅スペクトルは

$$S3 = S1 - \alpha * S2$$

で求める．この α の値は各音源について推定する必要がある．α の値が大きすぎれば，伴奏を引いた結果負になる数が増える，この負になった数の割合の閾値を超えない最大値を α とする．

3.1.3　F0 推定

伴奏を引いた振幅スペクトルを再合成した音源から F0 を推定する．この手法で得られる無伴奏歌唱音源は音声解析のため録音された音源とは異なるため，耐雑音性に優れたものが望ましい [1]．そのため PYIN[2] や LimingShi 法 [3] 等の推定方法が考えられる．本研究では PYIN を用いた．

3.2　音高分布の可視化

音高（F0）の可視化のための画像処理的なアプローチ [4] について，処理手順に沿って論じる．処理の詳細に関しては [4] を参考にされたい．

3.2.1　音高データの表記

本章では歌唱者集合 S を構成する各歌唱者の音高の推移を以下のように表記する．

$$S = \{s_1, s_2, ..., s_n\}$$
$$s_i = \{p_{i1}, p_{i2}, ...p_{im}\} \quad (1)$$

ここで s_i は i 番目の歌唱者による歌唱の音高系列，n は歌唱者の総数，p_{ij} は i 番目の歌唱者の j 番目の時刻における F0 値の対数である．

3.2.2　グレースケール画像の生成

本手法では，時刻を横軸，周波数の対数を縦軸とした長方形領域を設定し，これを格子状に分割する．式 1 に示す p_{ij} の各々が上述の格子構造のいずれの長方形領域に該当するかを算出する．以上の処理による集計結果は 2 次元ヒストグラムを構成するが，本手法ではこれを横 N 画素，縦 M 画素の画像として扱う．長

方形領域に包括される p_{ij} の個数を集計した変数 r_{uv} から，以下の式

$$I_{uv} = 1.0 - (\alpha r_{uv})^{\gamma} \quad (2)$$

によって，左から u 画素目，下から v 画素目の明度 I_{uv} を求める．ここでの α および γ はユーザが調節可能な変数とする．

3.2.3 ラベリングによる頻出 F0 推移領域の特定

頻出する F0 推移を見つけやすくするための一手段として，本手法では上述の画像を閾値 β によって白黒 2 値化し，さらにラベリング処理を適用する．2 値化された画像の左から u 画素目，下から v 画素目の画素値 $B_{uv} = 1$ である画素を 1 個抽出し，隣接画素で $B_{uv} = 1$ であるものを再帰的に探索する．そして，探索が終了するまでに訪問した画素の集合に固有のラベルを割り当てる．この処理を $B_{uv} = 1$ である全ての画素に割り当てる．閾値 β はユーザが調節可能な変数である．

4 実行例

本手法による可視化の例を紹介する．プログラミング環境は Java 1.10.0 および JOGL (Java OpenGL) 2.3.2 を用い，実行例には DAMP-balanced dataset [1] に収録された”Let It Go”の 2024 人の歌唱を用いた．本研究では音響データから STRAIGHT[5] を用いて推定した F0 を入力とした．可視化結果の画素数は $N = 1000, M = 500$ とした．

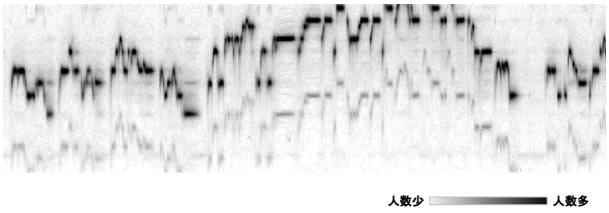

図 1: グレースケール画像として表示した例

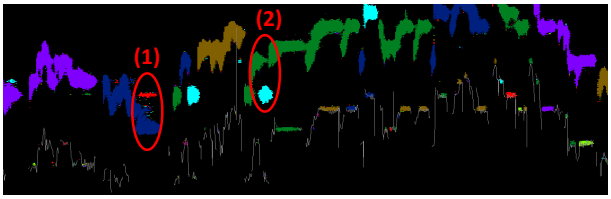

図 2: ラベリングした例

図 1 は周波数推移分布をグレースケール画像として表示した例である．黒に近いほど多くの歌唱者が同じ音高推移をとっていることを示している．画像中の上部に黒に近い部位が左右に分布しており，多くの歌唱者が同様な音高をとっていることがわかる．さらに，これと同様な動きが画像中の下部にもうっすらと見えるがこれは少数の歌唱者が音高を 1 オクターブ低く歌唱していた可能性が考えられる．

図 2 は図 1 にラベリングを適用した例である．この結果では，多くの歌唱者が有する同様な周波数推移に固有の色が割り当てられている．なお，この図ではマウスオーバーした箇所に該当する 1 人の歌唱者の周波数推移を折れ線グラフとして同時に表示している．この図の楕円 (1) に着目すると，多くの人が同様な歌唱をしていることが紺のラベルで表示されているのに対して，高い周波数で赤のラベルが存在することがわかる．さらに楕円 (2) にも同様の特徴がみられる．この可視化結果から，一定数の歌唱者に同様な癖がある可能性が考えられる．

[1] https://ccrma.stanford.edu/damp/

5 まとめと今後の課題

本報告では，同一楽曲に対する多数の歌唱データに対して推定した音高の推移を時系列データとみなし，画像処理的なアプローチによって可視化する手法を提案した．さらに一般的な楽曲音源からのデータ生成手法について述べた．今後の課題として，再生数などの付随情報を可視化したり，原曲の歌唱を正解データとして可視化するなどの機能を実装したい．さらに別の問題として，現状の実装では，多数の歌唱者で頻出する音高推移パターンを発見することには向いているが，例外的な音高推移を発見するのが難しいという点もあげられる．この点についても，ラベリング処理から外れた音高推移を折れ線グラフや散布図で表示するという形での解決が考えられる．このように可視化手法を改良したのち，さらに多様なデータに対し本手法を適用したいと考えている．

6 謝辞

本研究を進めるにあたり，産業技術総合研究所中野 倫靖氏，深山 覚氏，濱崎 雅弘氏，後藤 真孝氏にはデータ提供および多大な助言を賜りました．ここに感謝申し上げます．

参考文献

[1] 森勢 将雅, 河原 英紀, 基本周波数推定法の性能を概観するフレームワークの試作, 情報処理学会研究報告音楽情報科学（MUS), 2016-MUS-110, 1, pp.1–6, 2016.

[2] M. Mauch, S. Dixon, PYIN: A fundamental frequency estimator using probabilistic threshold distribu-tions, Proc. ICASSP2014, pp. 659–663 , 2014.

[3] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, M. G. Christensen, Robust Bayesian Pitch Tracking Based on theHarmonic Model, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27, 11, pp.1737–1751, 2019.

[4] 伊藤 貴之, 中野 倫靖, 深山 覚, 濱崎 雅弘, 後藤 真孝, 同一楽曲に対する多数の歌唱の基本周波数推定値分布の可視化, 情報処理学会研究報告音楽情報科学 (MUS), 2019-MUS-123, 47, pp.1–6, 2019.

[5] 河原 英紀，森勢 将雅， TANDEM-STRAIGHT と音声モーフィング：感情音声と歌唱研究への応用, 日本音声学会 論文誌, 13, 1, pp. 29–39, 2009.