ハーモニック・クラスタリングによる多重音基本周波数抽出における音源数およびオクターブ位置の推定*

◎亀岡弘和　西本卓也　嵯峨山茂樹　（東大・情報理工）

1　はじめに

複数音源の音響信号が混在したものを多重音と呼ぶ．多重音の基本周波数抽出は，音楽情報科学の分野において重要な要素の一つとされている．柏野らは，人間の知覚に関する仮説と楽音や音楽構造の統計データによる仮説を統合した Bayesian ネットワークを構成し，事後確率を評価関数として音名，周波数成分を出力する手法を提案した [1]．また後藤は，単一音の倍音構造をモデル化した確率分布を混合し，重みに関する最尤推定を行う手法を提案した [2]．両手法は高い分解能で基本周波数が抽出できるという特徴があり，これは音源分離や，非音楽的な音 (会話音声など) の基本周波数抽出などにおいて重要となる．また両手法に限らず多くの手法では対象音源の倍音比情報を活用しているが，実際は事前に得られない場合や倍音比の仮定自体が困難な音源 (音声など) を対象とする場合もある．我々は，必ずしも音源の倍音比情報の仮定を必要とせずに単一フレーム・単一チャネルの短時間スペクトルから基本周波数を連続値として抽出する方法 “ハーモニック・クラスタリング” を提案し，これまでに実音楽信号を対象として収束性の動作確認を行った [4]．ここでは，音源数*とオクターブ位置に関する問題は扱っていなかったため，本報告では AIC[3] を導入した音源数およびオクターブ位置の推定方法を検討し，これらを統合した基本周波数抽出アルゴリズムの動作確認を行う．

2　Harmonic Clustering の定式化

短時間周波数解析による観測スペクトルは，窓関数や分析区間内におけるピッチ変化により周波数方向に広がり幅をもつ．そのようなスペクトルを微小エネルギーの度数分布と解釈し，これに応じて微小エネルギーのクラスタリングを行えば，1 つの周波数成分に対応するスペクトルの広がりの周波数幅をクラスタ帯域として扱える．単一音の場合，単純なクラスタリングではなく，複数のクラスタ重心が倍音構造をなすような拘束条件の下でクラスタリングを行うことで基本周波数を抽出できる．以後，倍音構造に関して倍音周波数は基本周波数の整数倍であると仮定する†．従って，対数周波数スケールでは，あるクラスタ重心を μ(基本重心と呼ぶ) とすると，その他のクラスタ重心は $\mu+\log 2,\cdots,\mu+\log n,\cdots$ となる．このような拘束条件をもつ複数のクラスタをクラスタ群と呼び，このクラスタリングを “ハーモニック・クラスタリング (Harmonic Clustering)” と呼ぶ．

2.1　多重音スペクトルのクラスタリング

複数 (K個) のクラスタ群を用いて上述したクラスタ群の決定問題を多重音に対し次の 2 点に従い同様に扱う．ただし，倍音クラスタ群 k の n 倍音に対応するクラスタ帯域を B_n^k とし，その重心を $\mu_k+\log n$ とする．また，クラスタ群 k において，上限 (ナイキスト周波数の対数) までとりうるクラスタ重心数を N_k とする．

(i) 重複周波数成分の分離を考慮し，クラスタ帯域を確率的な尺度により分割する．そこで，対数周波数軸座標 x にある微小エネルギーのクラスタ帯域 B_n^k への帰属確率を $p_n^k(x)$ とする．

(ii) 音源間のパワー比や周波数成分間のエネルギー比 (倍音比) を考慮し，クラスタごとの重み w_n^k を導入する．ただし，式 (1) を満たすものとする．

$$\sum_{k=1}^{K}\sum_{n=1}^{N_k} w_n^k = 1 \tag{1}$$

以上に従い，ξ と ε との距離を表す関数 $\varphi(\xi,\varepsilon)$ を用いてクラスタリング評価関数を以下とする．

$$D(\boldsymbol{\mu},\boldsymbol{w}) = \sum_{k=1}^{K}\sum_{n=1}^{N_k}\int_{-\infty}^{\infty} w_n^k\cdot\varphi(x,\mu_k+\log n)\cdot p_n^k(x)\cdot f(x)\,dx \tag{2}$$

観測スペクトルにおける広がりの形状を正規分布で近似すると，クラスタ帰属度 $p_n^k(x)$ は，平均 $\mu_k+\log n$，分散 σ の正規分布 $g(x|\mu_k+\log n,\sigma^2)$ と (ii) で導入した重み w_n^k を用いて以下のように置ける．

$$p_n^k(x) = \frac{w_n^k\cdot g(x|\mu_k+\log n,\sigma^2)}{\sum_k\sum_n w_n^k\cdot g(x|\mu_k+\log n,\sigma^2)} \tag{3}$$

さらに，距離関数 $\varphi(x,\mu_k+\log n)$ を，重み w_n^k をかけた正規分布 $g(x|\mu_k+\log n,\sigma^2)$ の対数尤度

$$\varphi(x,\mu_k+\log n) = \log w_n^k\cdot g(x|\mu_k+\log n,\sigma^2) \tag{4}$$

と置くことで，この評価関数は EM アルゴリズムによる混合正規分布のパラメータの最尤推定における Q 関数とほぼ同値となる‡．以上より，初期設定 (ステップ 0) を経て，以下のようなステップ 1 とステップ 2 の反復計算の収束性は保証され，評価関数または Q 関数を局所最大化する $\boldsymbol{\mu}$，$\boldsymbol{w}$ を得ることができる．

ステップ 0: (初期設定)
重心系列 $\boldsymbol{\mu}$，重み系列 $\boldsymbol{w}$ の初期値を与える．

ステップ 1: ($p_n^k(x)$ の更新，E ステップに相当)
式 (2) により $p_n^k(x)$ を算出し，$D(\boldsymbol{\mu},\boldsymbol{w})$ を求める．

ステップ 2: ($\boldsymbol{\mu}$ の更新，M ステップに相当)
$p_n^k(x)$ を固定の下で，以下により $\boldsymbol{\mu}$，$\boldsymbol{w}$ を更新後，ステップ 1 に戻る．

$$\{\boldsymbol{\mu},\boldsymbol{w}\} = \underset{\{\bar{\boldsymbol{\mu}},\bar{\boldsymbol{w}}\}}{\operatorname{argmax}} D(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{w}}) \tag{5}$$

混合正規分布モデル (以後，単純にモデルと呼ぶ) が $f(x)$ に従って分布する微小エネルギーを生起する対数尤度はステップ 2 により単調増加でき，反復計算により局所最大化できる．以上の手順では，音源数が K 個であるという前提の下で，K 個の局所最尤のピッチクラス§を抽出することができる．そこで，次章では音源数とオクターブ位置方法について述べる．

3　音源数とオクターブ位置推定

3.1　処理行程 1: ピッチクラスと音源数推定

最大対数尤度は，真の音源数に関わらずクラスタ群数が多いモデルほど大きく出やすいという傾向があり，音源数を判定する尺度としては適当ではない．そこで，近似的に期待対数尤度の不偏推定量と同じ意味をもつ AIC(赤池情報量規準)[3] を導入し，最適なモデルの選択規準とする．同時発音されるピッチクラス数は 12 以下であると仮定し¶，12 個のクラスタ群から開始し，尤度に比較的関与しないと見なせる倍音クラスタ群を順次削減させながら，その都度，AIC を計算する．AIC が最小となるモデルを最適と考え，そのときのクラスタ群数を推定音源数とする．具体的な手順を以下に示す．

---

* “Estimation of Number of Sound Sources and Octave Positions in Multipitch Extraction using Harmonic Clustering” by Hirokazu Kameoka, Takuya Nishimoto and Shigeki Sagayama (Graduate School of Information Science and Technology, The University of Tokyo).

*本報告で言う音源数とは，同時発音する基本周波数の数をさす．

†ベルなどのように非調和性の著しい楽器音はここでは扱わないが，整数倍の代わりに対象とする音源の倍音構造に合わせた拘束を与えれば同様に扱える．

‡$f(x)$ が確率密度関数ではなくスペクトル密度関数である．

§オクターブ位置を区別せずに分類された音高の集合

¶12 平均律の音楽においては，12 種類のピッチクラスが基本音階であるため妥当な仮定といえる

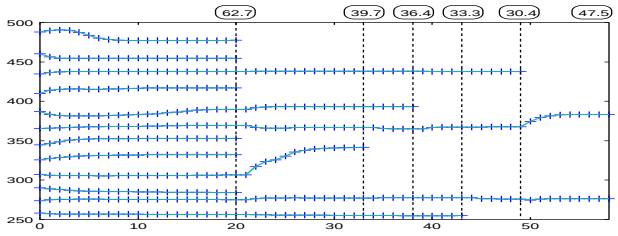

図 1: クラスタ群数および基本重心の更新

1. あるオクターブ範囲内における 12 平均律音階の基本周波数を基本重心の初期値とする．
2. 2.1 の反復計算により最尤パラメータを求める．ただし，ここではクラスタごとの重み w_n^k に関して
$$w_1^k = w_2^k = \cdots = w_{N_k}^k \ (= w^k) \tag{6}$$
のような拘束を与え，クラスタごとではなくクラスタ群ごとの重みを用いる．この場合，ステップ 2 における μ_k，w^k の更新値は以下で与えられる．ただし，$F{=}\int_{-\infty}^{\infty} f(x)dx$ とする．
$$\bar{\mu}_k = \frac{\sum_{n=1}^{N_k}\int_{-\infty}^{\infty}(x-\log n)p_n^k(x)f(x)dx}{\sum_{n=1}^{N_k}\int_{\infty}^{\infty}p_n^k(x)f(x)dx} \tag{7}$$
$$\bar{w}^k = \frac{1}{FN_k}\sum_{n=1}^{N_k}\int_{-\infty}^{\infty} p_n^k(x)\ dx \tag{8}$$
3. AIC を算出する．クラスタ群ごとに 2 つのパラメータ μ_k, w^k があるので，自由パラメータ総数は$2{\times}K$である．AIC が上昇した時点で終了し，削減前のクラスタ群数$\hat{K}$ を推定音源数とする．
4. 以下によりクラスタ群を削減し，残ったクラスタ群数を K とする．$K{=}\hat{K}$とし，2. に戻る．
   (a) w^k が最小のクラスタ群を消滅させる．これは，モデルが与える最大対数尤度への関与度が最も低いと見なせるためである．
   (b)2 つの基本重心間の距離が $(1/12)\log 2$ より小さいとき，重みの小さい方のクラスタ群を消滅させる．これは，1 つの極値に 2 つの基本重心が収束していると考えられるためである．

この行程をあるスペクトルに対して実際に行った例を図 1 に示す．破線が 2. において$\boldsymbol{\mu}$と$\boldsymbol{w}$が収束したと見なされた時点を表し，丸で囲んだ数値は 3. において計算したAICの値である．クラスタ群の数が 2 のときにAICが初めて上昇したため，推定音源数は 3 となる．

3.2 処理行程 2: オクターブ位置推定

3.1 で求まった基本重心は，真の基本周波数と同一のピッチクラスに対応するが，オクターブ位置も同一であるとは限らない．従って，基本重心をさまざまなオクターブ位置に置いた場合のモデルがすべて候補となる．最大対数尤度は，真のオクターブ位置に関わらず基本重心のオクターブ位置が低いモデルほど大きく出やすいという傾向があり，オクターブ位置を判定する尺度としては適当ではない．従って前節同様，モデルの選択規準として AIC を導入する．ここでは，オクターブ位置の異なる同一ピッチクラスが同時発音した音 (以後，オクターブユニゾンと呼ぶ) に関しては，低い方の基本周波数による単一音と見なす．行程 1 により得られる基本クラスタ重心 μ_k に対し，オクターブを低い位置から順に上げていき，その都度，AIC を計算する．AIC が最小となるモデルを最適と考え，そのときのオクターブ位置を推定位置とする．オクターブ位置を表す指標 t (自然数) を用いて，以下に示す手順を $\hat{K}$ 個のクラスタ群すべてについて行う．

1. 基本重心を $\mu_k + \log t$ と置き (t の初期値は 1)，上限までとりうるクラスタ重心の数を N_k^t とする．
2. 2.1 の反復計算により最尤パラメータを求める．ステップ 2 における w_n^k の更新値は以下である．基本重心は固定とする．

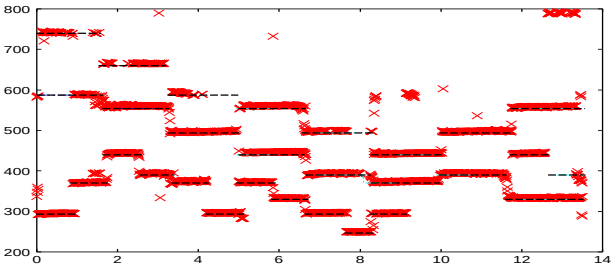

図 2: “Kanon” の一部分における抽出基本周波数

$$\bar{w}_n^k = \frac{1}{F}\int_{-\infty}^{\infty} p_n^k(x)\ dx \tag{9}$$

3. AIC を算出する．このとき，自由パラメータ総数は N_k^t である．AIC が上昇した時点で終了し，上昇前における t を推定オクターブ位置の指標とする．t を 1 増やし，1. に戻る．

4 動作実験

2 つの処理行程を統合した基本周波数抽出アルゴリズムの動作実験を行った．サンプリング周波数を44.1kHz，フレーム長を25ms，フレームシフトを10ms とし，Hamming 窓を窓関数として周波数解析 (FFT) を行った．実験データとしては，J. Pachelbel 作 “Kanon” のヴァイオリン演奏家による三重奏の実録音信号を用いた．

基本周波数の抽出結果の一部を図 2 に示す．×印がフレームごとに抽出した基本周波数を，破線が正解音名に対応する基本周波数を表す．行程 1 における音源数正解率は 94.9%，両行程を通した音名正解率は 92.7%であった．ただし，オクターブユニゾンに関しては音源数は 1，オクターブ位置は低い方を正解とした．正解率は，音高の変化時のフレームを目視で調べ，全フレームに対する正解フレーム数の割合とした．行程1と2ではAIC を最小にするモデルを必ずしも選択できるという保証はないが，結果を見る限り妥当な方法であったといえる．また，AIC が音源数とオクターブ位置推定に有効に活用できることが確認できた．

5 おわりに

本報告では，ハーモニック・クラスタリングとAIC の導入による音源数推定法とオクターブ位置推定法を統合した，単一チャネルおよび単一フレーム処理によって多重音の基本周波数を抽出するアルゴリズムを提案した．モノラル音楽信号を対象として動作確認を行い，推定に AIC が有効利用できることを確認した．今回オクターブユニゾンを単一音と見なしたが，倍音比情報を一切用いずに複数音として見分けることは原理的に不可能である．従って，事前の倍音比情報を必要としない提案手法の利点を保持した方法論を追求していく場合には，倍音比情報を自動獲得する方法などの検討を行う必要がある．また，現段階では単一フレームごとの処理として徹底して取り組んできたが，複数フレームにまたがった時間方向のグルーピングを行った方が性能の上昇だけではなくオクターブユニゾン問題の解決も期待できるため，時間方向も含めた 2 次元クラスタリングへの拡張についても検討していきたい．

参考文献

[1] 柏野邦夫, 木下智義, 中臺一博, 田中英彦: “音楽情景分析の処理モデル OPTIMA における和音 の認識,” 電子情報通信学会論文誌, Vol. J79-D-II, No. 11, pp. 1762–1770, 1996.

[2] M. Goto: “A Predominent-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models,” *Proceedings of the 2001 IEEE International Conference on Acoustics*, Speech and Signal Processing, ICASSP 2001, pp. V-3365–3368, 2001.

[3] H. Akaike: “Information Theory and an Extension of the Maximum Likelihood Principle,” *2nd Inter. Symp. on Information Theory*, Akademia Kiado, Budapest, pp. 267–281, 1973.

[4] 亀岡弘和, 西本卓也, 篠田浩一, 嵯峨山茂樹: “ハーモニッククラスタリングによる多重音の基本周波数推定アルゴリズム,” 情報処理学会研究報告, SIGMUS50-5, pp. 37–43, 2003.