Harmonic-GMM の最尤推定と情報量規準に基づく多重音の基本周波数検出および調波構造分離*

亀岡弘和　西本卓也　嵯峨山茂樹
東京大学 情報理工学系研究科

1 はじめに

多重音から基本周波数 (以後 F_0 と呼ぶ) を検出する技術は，様々な貢献が期待される．例えば，Audio CD からの MIDI(着メロやカラオケ伴奏 etc.) 変換や採譜などの自動化および支援，または音楽・音声圧縮符号化効率や音声認識の精度の向上などが挙げられる．しかし，短時間周波数解析によるスペクトルの広がり，調波成分の重複，ミッシングファンダメンタルなどの複合的な要因により，容易な問題ではない．

Chazan らは時間伸縮波形モデルの最適近似と櫛形フィルタにより同時発話音声から音声を分離する手法を提案した [1]．Klapuri はフィルタバンク処理によるロバストな F_0 検出を実現した [2]．後藤は混合正規分布による多数の調波構造モデルの各重みを最大事後確率推定により求め，それに基づいて F_0 を追跡する手法を提案した [3]．

上の例のようにこれまで様々な定式化により F_0 検出手法が提案されてきたが，発音数の推定や「倍ピッチ/半ピッチエラー」の問題を厳密に定式化した手法はいまだ報告されていない．本報告では，多重音スペクトルの解析を統計的推定手法に帰着させ，情報量規準に基づいて同時発音数，真の F_0 を適切に推定する新しいアルゴリズムを導く．また，これは各音ごとに調波構造を分離する特徴も併せもつ．

2 Harmonic-GMM の定式化

短時間周波数解析では，一般に解析区間に窓関数を掛けるため，左右に広がりをもつスペクトルが観測される．このため，低い周波数分解能のスペクトルから周波数を容易に求められない．また，複数の音の調波成分が重複すれば，検出はさらに困難になる．

窓関数として正規分布窓を用いれば，スペクトルの広がりの形状が理論的に正規分布の形状となるので，基本周波数成分に対応する正規分布の平均だけが自由度をもつ拘束つきの混合正規分布 (GMM, Gaussian Mixture Model) により単一音の調波構造をモデル化できる．これを Harmonic-GMM と呼ぶ．k 番目の Harmonic-GMM(G_k) の各平均は，$\boldsymbol{\mu}_k = \{\mu_k, 2\mu_k, \cdots, n\mu_k, \cdots, N_k\mu_k\}$ と書ける．ただし，n は n 次高調波成分に対応する正規分布のラベルを，N_k は正規分布の数を表す．

K 個の音の調波構造が重なり合うスペクトルを，Harmonic-GMM を K 個混合することによりモデル化し，モデルパラメータを，$\{\theta\} = \{n\mu_k, w_n^k, \sigma_n^k \mid k = 1, \cdots, K\}$ とする．w_n^k，σ_n^k は n 次成分の重み，分散を表す．スペクトル分布を正規化して確率変数 (周波数)ω の確率分布 $f(\omega)$ と見なす．θ の事後確率を最大化する θ は，多重音モデルの ω における平均対数尤度 $f(\omega)\log p(\omega|\theta)$ を用いて以下で表される．

$$\theta = \operatorname*{argmax}_{\theta} \left\{ \log p(\theta) + \int_{-\infty}^{\infty} f(\omega) \log p(\omega|\theta) \mathrm{d}\omega \right\} \quad (1)$$

$p(\theta)$ は θ の事前確率を表し，これを一様分布とすれば，右辺第二項を最大化すること (最尤推定) と等価になる．式 (1) を解析的に解くことは困難であるが，EM(Expectation Maximization) アルゴリズムにより以下の Q 関数を最大化する $\bar{\theta}$ を θ の更新値として逐次的に計算することで局所最適解を得ることができる．

$$Q(\theta, \bar{\theta}) = \log p(\bar{\theta}) + \sum_{k=1}^{K} \sum_{n=1}^{N_k} \int_{-\infty}^{\infty} p(n,k|\omega,\theta) f(\omega) \log p(\omega,n,k|\bar{\theta}) \mathrm{d}\omega \quad (2)$$

3 多重音 F_0 検出アルゴリズム

3.1 発音数推定プロセス

EM アルゴリズムにより得られるモデルパラメータの収束値は初期値に依存し，しばしば誤った局所解に陥る．そこで，予想される発音数より多めの数の Harmonic-GMM を満遍なく初期配置すれば μ_k の目的解を得る可能性は高くなるはずである．ただし，Harmonic-GMM は発音数と同数あれば十分であり，この場合最尤の多重音モデルは観測スペクトルに対して明らかに過適応を起こしている．ここで，情報量規準の一つとしてよく知られる[1]赤池情報量規準 (AIC, Akaike Information Criterion)[4] を導入し，適切な自由パラメータ数を決定することで発音数推定を試みる．すなわち，不必要な Harmonic-GMM(後述) から削減していき，AIC が最小となるときの数を推定発音数と考える．具体的な処理手順を以下に示す．

1. 限定した周波数帯域内に基本平均を K 個配置する．
2. EM アルゴリズムにより θ の最尤推定値を求める．ただし，ここでは正規分布の重みは平均と同様 k のみに依存するパラメータ w_k とし，G_k ごとの重みを表す．調波成分の強度比を事前にモデルに与えることも可能である．式 (2) を最大化する μ_k，w^k，σ_n^k の更新値は偏微分を 0 と置くことで以下として求まる．

$$\bar{\mu}_k = \frac{\sum_{n=1}^{N_k} \int_{-\infty}^{\infty} p(n,k|\omega,\theta) f(\omega) n\omega \ \mathrm{d}\omega}{\sum_{n=1}^{N_k} \int_{-\infty}^{\infty} p(n,k|\omega,\theta) f(\omega) n^2 \ \mathrm{d}\omega} \quad (3)$$

$$\bar{w}^k = \sum_{n=1}^{N_k} \int_{-\infty}^{\infty} p(n,k|\omega,\theta) \ \mathrm{d}\omega \quad (4)$$

$$\bar{\sigma}_n^k = \sqrt{\frac{\int_{-\infty}^{\infty} p(n,k|\omega,\theta) f(\omega)(\omega - n\mu_k)^2 \mathrm{d}\omega}{\int_{-\infty}^{\infty} p(n,k|\omega,\theta) f(\omega) \mathrm{d}\omega}} \quad (5)$$

3. AIC を算出する．AIC が最小値をとるときの Harmonic-GMM の数 $\hat{K}$ を推定発音数とする．
4. w_k が最小の (尤度への関与が最も低く，不必要と見なせる) Harmonic-GMM を削除する．分散 σ_n^k を大きめの値に[2]置き換え，ステップ 2 に戻る．

3.2 F_0 と調波スペクトル成分検出プロセス

前プロセスにおいて求まる μ_k の局所最適解は，真の F_0 だけではなくその整数倍あるいは整数分の 1 倍のいずれかに対応する可能性がある．ある音の調波成分と他の音の調波成分が完全に重複する多重音の場合は単一音と見なし，ここでは各調波成分の強度を手がかりとして真の F_0 を検出する．μ_k を整数倍/整数分の 1 倍に置き換えながら，その都度正規分布ご

---

* "Fundamental Frequency Detection and Spectral Separation of Mixed Sound Based on Harmonic-GMM and Information Criterion" by Hirokazu KAMEOKA, Takuya NISHIMOTO, and Shigeki SAGAYAMA (The University of Tokyo).

[1] AIC は $-2\times$(最大対数尤度)$+2\times$(自由パラメータ数) で与えられる．

[2] σ_n^k の更新は，分散の推定値を得るためではなく，大きい初期値を与えることで μ_k の目的解への収束を促進するのが狙いである．

との重み w_n^k の最尤推定値から調波成分の強度比を推定する．もし，置き換えた μ_k が真の F_0 の整数分の1倍である場合，実際に存在する調波成分に対応する重み以外は全体のモデルが与える平均対数尤度にほとんど関与しないはずであり，モデルは過適応を起こしていると言える．この観点から，前説同様 AIC に基づいて真の F_0 を検出することを試みる．前プロセスにおいて残った Harmonic-GMM すべてについて以下を行う．

1. G_k の 1 次成分の平均を $t\mu_k$ に置き換える．ただし，t を初期値 1 の自然数とする．限定した周波数帯域内まででとり得る正規分布の数を N_k^t とする．
2. EM アルゴリズムにより w_n^k の最尤推定値を求める．式 (2) を最大化する w_n^k の更新値は以下となる．

$$\bar{w}_n^k = \int_{-\infty}^{\infty} p(n,k|\omega,\theta)\ \mathrm{d}\omega \tag{6}$$

3. 自由パラメータ数を N_k^t として AIC を算出する．t を 1 増やし，ステップ 1 に戻る．AIC が最小となる $t\mu_k$ が推定 F_0 となる．また，最終的な w_n^k の最尤推定値が各音の調波成分強度比の推定値となる．

4 評価実験

複数話者による同時発話音声と音楽音響信号を対象として F_0 検出アルゴリズムの評価実験を行った．

音声と音楽の各信号はサンプリング周波数 12kHz と 44.1kHz，フレーム長 46ms，フレームシフト 10ms とし，正規分布窓を窓関数として FFT によりスペクトル系列を得た．音声信号とフレームごとの正解の F_0 値は ATR 音声データベース A セットのデータを用いた．音楽信号は，RWC 研究用音楽データベース，CD に収録されている 2 曲と実演奏を採録した 2 曲を用い，F_0 の正解はスペクトル系列からの目視と楽譜を手がかりに手作業で[3]ラベリングした．ラベリングされた正解 F_0 値から 5%の誤差範囲内の値で F_0 が検出できた場合に正解と見なし，各音に関して正解した延べフレーム数を割り出し，その音が発音している延べフレーム数に対する割合を正解率とした．

各実験データに対する F_0 検出の結果を表 1，表 2 に示す．ミッシングファンダメンタルが見られた場合でも正しく推定ができ，音声信号，音楽信号に対して 85%前後の正解率を得た．AIC がある程度有効利用できることが確認できたが，相対的に強度の小さい音が無視されてしまう傾向が見られ，発音数の正解率に反映された (表 2)．人間が知覚する音量は対数パワーに比例することが知られているが，モデル選択が線形スペクトルに基づいていることが大きな原因として考えられる．

5 調波成分の強度比パラメータの導入

4 章では，絶対協和和音などのように単一音と同等なスペクトル構造をなす和音は単一音として推定された．このような和音は，一部の調波成分だけが強め合うので，極端な音色の楽音を除けば起伏のあるスペクトル包絡が予想される．ここでは，これを複数音として推定する方法を予備的に検討する．

音 k の n 次調波成分と基本周波数成分との強度比を $r_n^k(r_1^k=1)$ とし，音ごとの強度比を w_k とすれば，モデルの各正規分布の重み w_n^k は $r_n^k \cdot w_k$ と表され，これを用いて 3 章と同様の操作を行う．前半プロセスでは，$1/f$ ゆらぎを考慮して $r_n^k = 1/n$(固定) と置き，後半プロセスでは r_n^k の事前分布 $p(r_n^k)$ を平均 $1/n$，分散 ν の正規分布と置くことで，最大事後確率推定により起伏の激しいスペクトル包絡を敬遠する制約の下で調波成分強度の推定を行うことができる．これにより，一つの Harmonic-GMM では補えない調波成分をオクターブ位置の異なる Harmonic-GMM が補う形となり，個々の F_0 が検出できることが期待される．このとき，EM アルゴリズムにおける w_k と r_n^k の更新値はそれぞれ式 (4)，式 (7) となる．

$$\bar{r}_n^k = \left(\frac{1}{n} + \sqrt{\frac{1}{n^2} + 4\nu^2 \int_{-\infty}^{\infty} p(n,k|\omega,\theta) f(\omega) \mathrm{d}\omega}\right) \Big/ 2 \tag{7}$$

ピアノ演奏による音楽信号データ 2 曲を対象として動作実験を行った．その結果を表 3 に示す．ここでは，絶対協和和音などは複数音と見なして評価した．

F_0 がおよそ 200Hz 以下のときのピアノ音のスペクトル包絡は周波数に反比例した形状から程遠く，低音域が和音に含まれている場合，上で立てた調波成分の強度比の仮定では，うまく機能しなかった．ただし，仮定した形状と包絡形状がある程度近い音高だけからなる和音に関しては，オクターブ奏法などの F_0 検出も適切に行えることが確認できた．

表 1: 同時発話音声を対象とした F_0 検出結果

実験データ (話者)	F_0 正解率 (%)
女性話者 2 人	86.4
男性話者 2 人	86.6
女性話者 1 人と男性話者 1 人	82.7

表 2: 音楽音響信号の発音数推定/F_0 検出結果 1

実験データ (作曲者/タイトル)	正解率 (%)	
	発音数	F_0
J.S.Bach: Ricercare à 6 (一部)	87.8	87.7
Komponist unbekannt: BWV Anhang 116	80.4	84.2
J.S.Bach: BWV 1046, no. 1, mov. 4	86.5	89.2
Pachelbel, Kanon	94.2	92.7

表 3: 音楽音響信号の発音数推定/F_0 検出結果 2

実験データ (作曲者/タイトル)	正解率 (%)	
	発音数	F_0
W.A.Mozart: Menuett, C dur	88.5	90.2
Komponist unbekannt: BWV Anhang 116	80.7	85.4

6 まとめ

本報告では，調波構造を GMM でモデル化し，情報量規準に基づいて同時発音数，F_0，調波成分強度比を推定する手法を提案した．また，モデルに調波成分の制約つき強度比パラメータを導入することで，絶対協和和音などから個々の F_0 が検出可能な拡張方法を検討した．音声信号および音楽信号に対して全体として 80%〜90%程度の正解率を得た．

今後は，F_0 の時間推移を考慮に入れた 2 次元の GMM に拡張させ，性能の向上を図る予定である．さらに，楽器音のスペクトル包絡モデルとその学習方法を現在考案中であり，いずれ本手法に導入させたい．また，大域的最適解を得やすくするよう改良された DA(Deterministic Annealing)EM アルゴリズムや，最大事後確率推定とモデル選択を同一の枠組として備える変分ベイズ法などの適用も検討したい．

参考文献

[1] D. Chazan, Y. Stettiner and D. Malah, "Optimal Multi-pitch Estimation Using the EM Algorithm for Co-channel Speech Separation," *Proc. ICASSP93*, Vol. 2, pp. 728–731, 1993.

[2] A. Klapuri, T. Virtanen and J. Holm, "Robust Multipitch Estimation for the Analysis and Manipulation of Polyphonic Musical Signals," *In Proc. COST-G6 Conference on Digital Audio Effects*, pp. 233–236, 2000.

[3] M. Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models," *Proc. ICASSP2001*, Vol. 5, pp. 3365–3368, Sep 2001.

[4] 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.

[3]なお，絶対協和和音などのように単一音と同等なスペクトル構造をなす和音は単一音と見なし，最も低い音高を正解とした．