拘束つき混合正規分布モデルの MAP 推定による同時発話音声の F_0 追跡*

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

複数の音が混在する多重音の単一チャネル信号から基本周波数 (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，会議や討論の場を想定した同時複数音声認識，電話のベルやテレビなどの妨害音に頑健な音声認識や，複数話者からの韻律情報抽出などが挙げられる．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現象などの複合的な要因により，容易に解決できる問題ではない．

音声信号処理の分野において近年この研究は徐々に盛んになり，いくつかの有効な手法が報告されている．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]．Wu らは，フィルタバンク処理と F_0 の勾配を状態とした隠れマルコフモデルを用いた F_0 追跡手法を提案した[2]．

上の例も含め，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，話者数の推定や “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，同時発話音声による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話者数，真の F_0 を適切に推定する “検出処理” と，直前フレームの検出結果に基づいて F_0 を追跡していく “追跡処理” により構成される新しいアルゴリズムを提案する．

2 拘束つき混合正規分布モデルの定式化

短時間周波数解析における窓関数や，解析区間内での周波数の連続的な変化などの影響により，左右に広がりをもつスペクトルが観測される．短時間分析ではスペクトルの周波数分解能は低いため，ローカルピークが必ずしも正確な周波数と一致しない．そこで，ピークを検出するという考え方から離れ，スペクトルの広がりの形状を正規分布で最適近似し，その平均を推定することで周波数検出を行うことを考える．

窓関数として正規分布窓を用いれば，窓関数の影響のみによるスペクトルの広がりの形状は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により単一音の調波構造をモデル化できる．これを調波モデルと呼ぶ．調波モデル 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 個の音の多重音スペクトルを，調波モデルを 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，$\{\theta\} = \{n\mu_k, w_n^k, \sigma_n^k \mid n,k \in \mathbf{N}\}$ 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 $f(\omega)$ と見なせば，θ の事後確率を最大化する θ は以下で表される．

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

$p(\theta)$ は θ の事前確率を表す．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 θ の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 同時発話音声の F_0 追跡アルゴリズム

本章では，話者数および各 F_0 を検出する処理 (“検出処理”) と直前フレームにおいて検出された F_0 に基づき F_0 を追跡する処理 (“追跡処理”) のいずれか一方をフレーム毎に実行し，逐次的に複数の F_0 を同時検出していくアルゴリズムについて述べる．

3.1 検出処理

発話開始時，フレーズ境界や新たな話者の音声介入時などにおいては，話者数とそれぞれの F_0 を検出する必要がある．この “検出処理” は，話者数推定ステップと F_0 検出ステップから成る．

話者数推定ステップ

まず，μ_k が目的解へ局所収束する可能性を高くするため，予想される発音数より多めの調波モデルを満遍なく初期配置する．ただし，調波モデルは話者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 (Akaike Information Criterion: AIC)[3] を導入し，適切な自由パラメータ数を推定する．すなわち，不必要な調波モデル (後述) から削減していき，AIC が最小となるときの数を推定話者数と考える．具体的な手順を以下に示す．

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 (事前分布を一様分布とする)．ここでは正規分布の重みは k のみに依存する調波モデルごとの重みパラメータ w_k とする．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．

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

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

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

3. AIC を算出する．AIC が最小値をとるときの調波モデルの数 $\hat{K}$ を推定話者数とする．
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) 調波モデルを削除する．分散 σ_n^k を大きめの値に[2]置き換え，2. に戻る．

---

*“F_0 Tracking of Simultaneous Speech Based on Maximum A Posteriori Estimation for Constrained Gaussian Mixture Model” by Hirokazu KAMEOKA, Takuya NISHIMOTO, and Shigeki SAGAYAMA (Graduate School of Information Science and Technology, 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．

F_0 検出ステップ

前ステップにおいて求まる 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．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 AIC に基づき，適切な μ_k の位置を推定する．前ステップにおいて残った調波モデルすべてについて以下を行う．

1. 調波モデルの 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, σ_n^k の最尤推定値を求める．M ステップにおける更新値は式 (6)，(5) 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，1. に戻る．AIC が最小となる $t\mu_k$ が推定 F_0 となる．

3.2 追跡処理

1 つのフレーズ区間では，ある時点の F_0 と直前の F_0 の間には強い依存関係があるはずである．そこで，直前フレームでの F_0 の検出結果を μ_k の事前分布に反映させ，最大事後確率 (Maximum A Posteriori, MAP) 推定により μ_k をフレーム毎に更新 (追跡) する．μ_k の (直前フレームでの μ_k の推定値に基づく) 予測値を $\hat{\mu}_k$ とし，μ_k の事前分布を $\hat{\mu}_k$ を平均，ν を分散とした正規分布とすれば，式 (2) より，EM アルゴリズムの M ステップにおける μ_k の更新値は

$$\bar{\mu}_k = \frac{\hat{\mu}_k + \nu^2 \sum_{n=1}^{N_k} \frac{n}{{\sigma_n^k}^2} \int_{-\infty}^{\infty} \omega p(n,k|\omega,\theta) f(\omega) \mathrm{d}\omega}{1 + \nu^2 \sum_{n=1}^{N_k} \frac{n^2}{{\sigma_n^k}^2} \int_{-\infty}^{\infty} p(n,k|\omega,\theta) f(\omega) \mathrm{d}\omega} \tag{7}$$

となる．また，重み w_n^k と分散 σ_n^k の更新はそれぞれ式 (6)，(5) を用いる．この追跡処理が連続で 3 フレーム以上続く場合，予測値 $\hat{\mu}_k$ は，過去の直前の 2 フレームにおける μ_k の推定値 μ'_k, μ''_k を結ぶ直線の延長上の値とし，$\hat{\mu}_k = 2\mu'_k - \mu''_k$ と定める．それ以外の場合は，直前フレームの推定値を予測値とする．

3.3 アルゴリズム構成

初期フレームは “検出処理” を実行し，以降のフレームでは，直前フレームでの多重音モデルと $f(\omega)$ の KL(Kullback-Leibler) 情報量が一定閾値以上の場合は “追跡処理” を，閾値より大きい場合は新たなフレーズの開始直後あるいは新たな話者による音声の介入直後と見なして改めて “検出処理” を行う．ただし，フレームごとの処理選択方法は今後改善の余地がある。なお，各フレームで求まる w_n^k の推定値をもとに話者ごとの音声信号を分離合成することもできる．

4 評価実験

上記の F_0 追跡アルゴリズムの性能を確認するため，F_0 検出方法としてよく知られる Cepstrum 法との比較を行った．また，すべてのフレームについて “検出処理” のみ行った場合 [4, 5] との比較も併せて行った．ATR 音声データベース A セットより，会話音声信号データ (サンプリング周波数 12kHz) とハンドラベルによる F_0 パターンの reference データを用いた．2 話者による同時発話音声信号を，SSR(signal-to-signal ratio)0dB で 2 つの音声信号を人工的に加算して作成した．各信号データに対し，フレーム長 64ms の正規分布窓をかけて周期 25ms で周波数解析 (FFT) を行い，スペクトル系列を得た．“検出処理” の話者数推定ステップにおいて，初期調波構造モデル数は 6 とし，μ_k を初期配置する周波数範囲は 90Hz から 360Hz とした．検出された F_0 が F_0 パターンの reference データの値から 5%以上外れた場合は gross error(それ以外は正解) と見なし，各話者ごとに正解した延べフレーム数をもとに正解率を算出した．

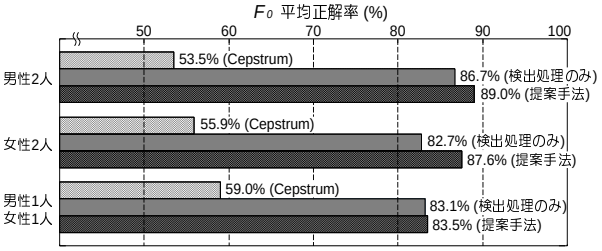

図 1: 各手法による F_0 検出の平均正解率

男性話者 2 人で作成したデータ 6 個，女性話者 2 人で作成したデータ 7 個，女性話者 1 人と男性話者 1 人で作成したデータ 8 個それぞれに対する各手法の平均正解率を図 1 に示す．提案手法の正解率は Cepstrum 法に比べて大きく上回り，基本性能の確認ができた．また，“検出処理” のみを行った場合に比べ，性能向上が確認できた．“検出処理” では情報量規準に基づいて話者数推定を行っているため，相対的に強度の小さい方の話者音声が無視されてしまう傾向があるが，“追跡処理” では調波モデルの削減はしないため，相対的に強度の小さい音声の F_0 の追跡もできたことが正解率向上の理由の 1 つとして考えられる．ただし，直前フレームの検出結果が誤っていた場合には，それ以降のフレームにも影響する危険性があり，“検出処理” の精度向上が今後の重要な課題である．

5 まとめ

本報告では，調波構造を混合正規分布でモデル化し，情報量規準に基づいて同時発話者数と F_0 を推定する “検出処理” と，直前フレームにおける推定結果をパラメータの事前分布として F_0 を推定する “追跡処理” により同時発話音声の F_0 追跡手法を提案した．

今後は，スペクトログラムを 2 次元の混合正規分布でモデル化し，追跡性能の向上を図る予定である．また，大域的最適解を得やすくするように改良された 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] M. Wu, D. Wang and G. J. Brown, “A Multi-pitch Tracking Algorithm for Noisy Speech,” *ICASSP2002*, Vol. 1, pp. 369–372, 2002.

[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, 嵯峨山 茂樹, “拘束つき混合正規分布の最尤推定と AIC による同時発話複数音声の F_0 推定,” 電子情報通信学会技術研究報告, Vol. 103, No. 519, SP2003-151, pp. 229–234, 2003.

[5] H. Kameoka, T. Nishimoto, S. Sagayama, “Multi-pitch Detection Algorithm Using Constrained Gaussian Mixture Model and Information Criterion for Simultaneous Speech,” *Proc. Speech Prosody 2004*, to be appeared, 2004.