社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

亀岡　弘和　　西本　卓也　　嵯峨山茂樹

東京大学大学院情報理工学系研究科
〒 113–0033 東京都文京区本郷 7–3–1
E-mail: {kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

あらまし　本報告では，複数話者による同時発話音声の単一チャネル信号に対する音声分離への拡張を念頭に置いた，混在する複数の基本周波数 (F_0) の推定アルゴリズムについて述べる．音声韻律 (F_0 パターン) の時間連続性は F_0 推定の際有用な情報であると考えられるが，今回は初期段階として各短時間分析窓それぞれ独立に処理を行うことを考える．まず，複数の調波構造が混在したスペクトルのモデルを，単一の調波構造をモデル化した拘束つき混合正規分布モデルを混合することで定式化する．このモデルのパラメータに関する最尤推定と情報量規準に基づくアルゴリズムにより，各分析窓において発話者数とそれぞれの F_0 およびスペクトル形状が検出できる．また，F_0 を連続値として高精度に推定できるという特徴をもつ．動作実験として話者一人による発話音声および話者二人による同時発話音声に対して Cepstrum 法との比較を行い，大きく上回る結果を得た．

キーワード　同時発話音声, 多重 F_0 推定, 混合正規分布, EM アルゴリズム, AIC

F_0 Contours Detection Based on Constrained Gaussian Mixture Model and Akaike Information Criterion for Simultaneous Utterances

Hirokazu KAMEOKA, Takuya NISHIMOTO, and Shigeki SAGAYAMA

Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
E-mail: {kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

Abstract　In this paper, a single-channel multi-pitch detection algorithm is described. Though temporal continuity of speech prosody (F_0 pattern) should be considered, we discuss a process done independently on each single frame as the first step. A model of multiple harmonic structures is constructed with a mixture of constrained Gaussian mixtures, with which a single harmonic structure is modeled. Our algorithm enables to detect both a number of concurrent speakers, and each spectral envelope of underlying harmonic structure based on a maximum likelihood estimation of the model parameters using EM algorithm and an information criterion. And it also extracts accurate F_0s as continuous values with simple procedures in spectral domain. Experiments showed our algorithm outperformed well-known cepstrum for both speech signals of a single speaker and simultaneous two speakers.

Key words　simultaneous speech, multi-pitch detection, Gaussian mixture model, EM algorithm, AIC

1. ま え が き

単一チャネル信号に対する音声分離の研究は，雑音重畳下の音声認識，会議や討論などの状況を想定した同時複数音声認識，音声強調，韻律分析，音声符号化や圧縮などに大きく貢献する．

単一チャネルでの音声分離のためには話者間の基本周波数 (以後 F_0 と呼ぶ) の違いに着目したアプローチをとることが一般に主流であり，信頼性の高い多重 F_0 推定手法が望まれる．音声分離を目的とした多重 F_0 推定手法においては，いかに同時発話者数を正確に推定できるか，あるいはいかに精度良く F_0 とスペクトル形状を抽出できるかが重要となり，かつ非常に難解な問題となる．これまで音響信号の分離を目的とした多重 F_0 推定手法は，音声信号処理 [1], [2] の分野だけでなく，音楽信号処理 [3]〜[5]，聴覚情景分析 [6]〜[8] の分野でも多数提案されて

きた．Chazan らは，長い分析窓において時間に伴って連続的に変化する F_0 に対して時間伸縮変換を最小二乗法により施すことで F_0 を一定にしたのち，櫛形フィルタを用いて音声分離を行う手法を提案した [1]．また Wu らは，フィルタバンク処理と，F_0 ダイナミクスを状態とした HMM (Hidden Markov Model) を用いた F_0 トラッキングによる雑音重畳化での多重 F_0 推定手法を提案した [2]．両手法は，精度の高い F_0 推定を実現し，良好な実験結果を得ているが，多くの従来法同様，同時発話者数を推定する具体的なプロセスを含んでいない．

我々は，同時発話者数とスペクトル形状が検出可能で，さらに F_0 を連続値として精度良く抽出できる多重 F_0 推定手法を実現することを目的とする．次章では提案するアルゴリズムの基盤をなす調波構造の混合モデルのパラメータ推定，3 章では同時発話者数とスペクトル形状の検出プロセスを含む多重 F_0 推定アルゴリズムについて述べ，4 章でアルゴリズムの動作実験の結果を報告する．

2. 拘束つき混合正規分布モデルの最尤推定

2.1 多重調波構造モデル

短時間スペクトルの解析では，解析区間に窓関数を掛けることが一般的である．そのため，周波数が一定の単一正弦波の信号であっても，線スペクトルではなく，左右に広がりをもつスペクトルが観測される．これは，窓関数のフーリエ変換と線スペクトルとの畳み込みを行うことに相当するためである．さらに分析窓区間内で周波数が連続的に変化する場合，それに応じた広がりをもつスペクトルが観測されることになる．これらに起因する基本周波数成分や調波成分の広がりにより，異なる信号同士の周波数成分が重なり合い，近接する周波数成分の分離や正確な F_0 あるいは高調波周波数の検出が困難となる(注1)．

そこで，このように広がって観測される周波数成分を各周波数の出現頻度分布あるいは確率分布と見なし，その分布を正規分布により近似することで，単一の調波構造を有するスペクトルを複数の正規分布の混合分布としてモデル化する．調和性の保持のため，基本周波数成分に対応する 1 つの正規分布の平均 (これを以後基本平均と呼ぶ) のみが自由度をもち，その位置に応じて残りのすべての正規分布の平均の位置は決定される．単一の調波構造をこのような拘束つきの混合正規分布によりモデル化したものを以後調波構造モデルと呼ぶことにする．基本平均を μ_k と置けば，調波構造モデル k の各平均 $\boldsymbol{\mu}_k$ は，対数周波数領域において，

$$\boldsymbol{\mu}_k = \{\mu_k, \mu_k + \log 2, \cdots, \mu_k + \log n, \cdots, \mu_k + \log N_k\} \quad (1)$$

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 は調波構造モデルごとの Nyquist 周波数まで取り得る正規分布の平均の数を表す．複数の調波構造が重なり合うスペクトルを，調波構造モデルをさらに混合することによりモデル化し，これを多重調波構造モデル $P_\theta(x)$ と呼ぶことにする．ただし，x は対数周波数とする．モデルパラメータ $\{\theta\}$ は，

$$\{\theta\} = \{\boldsymbol{\mu}_k, \boldsymbol{w}_k, \sigma \mid k = 1, \cdots, K\} \quad (2)$$

であり，$\boldsymbol{w}_k = \{w_1^k, \cdots, w_n^k, \cdots, w_{N_k}^k\}$ は調波構造モデル k の各正規分布の重み，σ は分散 (本報告では実験で定める定数とする)，K は混合された調波構造モデルの数をさす．

2.2 EM アルゴリズムによるモデルパラメータ推定

正規化した観測スペクトル $f(x)$ と上記した多重調波構造モデル $P_\theta(x)$ との Kullback-Leibler 情報量 $D(\theta)$ は以下となる．

$$D(\theta) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{P_\theta(x)} dx \quad (3)$$

$D(\theta)$ を最小化するモデルパラメータを求めることと，モデル $P_\theta(x)$ の平均対数尤度 ($f(x)$ を対数周波数 x の出現頻度を表す確率分布と解釈した場合，対数尤度の x に関する期待値) を最大化するモデルパラメータを求めることは等価である．そこで，x について，モデルパラメータ θ を $\bar{\theta}$ に更新したときのモデルの平均対数尤度の差は，

$$f(x)\log P_{\bar{\theta}}(x) - f(x)\log P_\theta(x) = f(x)\log \frac{P_{\bar{\theta}}(x)}{P_\theta(x)} \quad (4)$$

となる．Dempster らは，式 (4) において $f(x)$ を確率密度分布関数とし，平均対数尤度を最大にするために EM アルゴリズムを定式化した [9]．x がどの正規分布によって生成されたかは一意に決定できないため，これを直接最大化することはできない．そこで両辺に対し，x がどの正規分布から生成されたかを表す $P_\theta(n,k|x)$ についての期待値をとることで Q 関数と呼ぶ以下のような評価関数

$$Q(\theta, \bar{\theta}) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \int_{-\infty}^{\infty} P_\theta(n,k|x) f(x) \log P_{\bar{\theta}}(x,n,k) dx \quad (5)$$

を導出することができ，

$$\int_{-\infty}^{\infty} \left\{ f(x) \log P_{\bar{\theta}}(x) - f(x) \log P_\theta(x) \right\} dx \geqq Q(\theta, \bar{\theta}) - Q(\theta, \theta) \quad (6)$$

が成立するため，Q 関数を最大化する $\bar{\theta}$ を求め，逐次的に更新していくことで，x に関するモデルの平均対数尤度を単調に増加させることができる．$g(x|x_0, \sigma^2)$ を平均 x_0，分散 σ の正規分布とすると，$P_\theta(n,k|x)$ は，

$$P_\theta(n,k|x) = \frac{P_\theta(x,n,k)}{P_\theta(x)} \quad (7)$$

$$= \frac{w_n^k \cdot g(x|n\mu_k, \sigma^2)}{\sum_k \sum_n w_n^k \cdot g(x|n\mu_k, \sigma^2)} \quad (8)$$

$$g(x|x_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-x_0)^2}{2\sigma^2} \right\} \quad (9)$$

と書け，$\log P_{\bar{\theta}}(x,n,k)$ は各正規分布が与える対数尤度なので，以下となる．

(注1)：ミッシングファンダメンタル現象 (F_0 成分が極端に小さくなる現象) が生じたとき，F_0 を検出するために高調波周波数を正確に検出しなければならなくなる場合もある．

$$\log P_{\bar{\theta}}(x,n,k) = \log \frac{w_n^k}{\sqrt{2\pi\sigma^2}} - \frac{\{x-(\mu_k+\log n)\}^2}{2\sigma^2} \quad (10)$$

以上より，初期設定 (ステップ 0) を経て，以下のような E ステップ (*Expectation*-step) と M ステップ (*Maximization*-step) による反復計算の収束性は保証され，モデルの平均対数尤度を局所最大化するパラメータ $\boldsymbol{\mu}$，$\boldsymbol{w}$ を得ることができる．

ステップ 0: (初期設定)

モデルパラメータ $\boldsymbol{\mu}$，$\boldsymbol{w}$ の初期値を与える．

E-ステップ:

式 (5) により $Q(\theta,\theta)$ を計算する．

M-ステップ:

$Q(\theta,\bar{\theta})$ を最大化する $\bar{\theta}$ を計算する．

$$\theta = \underset{\bar{\theta}}{\operatorname{argmax}}\, Q(\theta,\bar{\theta}) \quad (11)$$

θ を更新後，E-ステップに戻る．

2.3 Clustering としての解釈

この拘束つき混合正規分布モデルの最尤推定は，スペクトル密度分布を架空の微小エネルギーの度数密度分布と捉えた場合，微小エネルギーを Clustering により各音へ分類する問題であると解釈することもできる．

正規分布の平均 $\mu_k + \log n$ をクラスタ n,k の中心と考えれば，事後確率 $P_\theta(n,k|x)$ を x に位置する微小エネルギーがクラスタ n,k に帰属する確率，対数尤度 $\log P_\theta(n,k,x)$ をクラスタ n,k の中心と x に位置する微小エネルギーとの距離を表す関数と見なすことができる．Clustering の評価関数は一般に，各微小エネルギーの位置と帰属するクラスタ中心との距離の自乗の和で表されるため，上記の観点により Q 関数と同一と見ることができる．ただし，Q 関数では事後確率 $P_\theta(n,k|x)$ と対数尤度 $\log P_\theta(n,k,x)$ は同一の確率分布に基づいて計算されるのに対し(注2)，Clustering においては必ずしもクラスタ帰属確率とクラスタ中心との距離関数は同一の関数を対応させる必要はない．例えば，微小エネルギーを最近傍のクラスタ中心のクラスタにすべて帰属させ(注3)，距離関数をユークリッド距離の自乗とした場合，帰属確率と距離関数は全く別の関数で表現される．このとき評価関数の最大化問題は，k-means アルゴリズムとして定式化できる．従って，上述したスペクトル成分に対する Clustering は，拘束つき混合正規分布モデルを EM アルゴリズムにより観測スペクトル分布の形状に近似する前節の定式化を包含する考え方であり，これを我々は “Harmonic Clustering” と呼んでいる [11], [12].

3. 多重 F_0 推定アルゴリズム

3.1 AIC によるモデルの選択基準

1 つのモデルに対する最尤パラメータは EM アルゴリズムにより求められるが，自由パラメータ数に応じてモデルの候補が複数個あるとき，その中から適切なモデルを選択する規準が必要となる．そこで，赤池によって提唱された AIC(Akaike Information Criterion，赤池情報量規準) [10] を導入する．AIC は

$$\text{AIC} = -2 \times (\text{モデルの最大対数尤度}) + 2 \times (\text{モデルの自由パラメータ数}) \quad (12)$$

で与えられ，適切な自由パラメータ数のモデルを選択する問題において有効であることが知られている．

3.2 同時発話者数検出プロセス

EM アルゴリズムにより得られるモデルパラメータの最尤推定量は初期値に依存し，しばしば誤った局所解に陥る場合がある．そこで，基本平均 μ_k の誤った局所解への収束を回避するため，予想される同時発話者数より多めの数の調波構造モデルを満遍なく初期配置しておくことで目的とする解が得られる可能性は高くなるはずである．ただし，このように初期配置された調波構造モデルの数が同時発話者数より多く，かつすべての目的解が得られているならば，多重調波構造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と言える．もし，周波数成分の分布の形状が正規分布で十分近似可能であれば，調波構造モデルは同時発話者数と同数あれば十分なはずである．そこで，不必要な調波構造モデルを順次削減していき，AIC が最小となる調波構造モデル数を判定することで同時発話者数を推定する．具体的な処理手順を以下に示す．

（1） 任意の周波数区間に基本平均 $\{\mu_1,\cdots,\mu_K\}$ の初期値を設定する．

（2） 2.2 節で述べた EM アルゴリズムにより最尤パラメータを求める．ただし，ここでは正規分布の重み w_n^k に関して

$$w_1^k = w_2^k = \cdots = w_{N(k)}^k (\ = w^k) \quad (13)$$

のような拘束を与える．これは，正規分布ごとではなく調波構造モデルごとの重みを規準として，優先的に削減すべき調波構造モデルを決定するためである (後述)．この場合，M-ステップにおける μ_k，w^k の更新値は式 (5) のそれぞれのパラメータに関する偏微分を 0 と置くことで得られ，以下で与えられる．

$$\bar{\mu}_k = \frac{\displaystyle\sum_{n=1}^{N_k}\int_{-\infty}^{\infty}(x-\log n)p_n^k(x)f(x)\,dx}{\displaystyle\sum_{n=1}^{N_k}\int_{\infty}^{\infty}p_n^k(x)f(x)\,dx} \quad (14)$$

$$\bar{w}^k = \frac{1}{N_k}\sum_{n=1}^{N_k}\int_{-\infty}^{\infty}p_n^k(x)\,dx \quad (15)$$

（3） AIC を算出する．調波構造モデルごとに 2 つの自由パラメータ 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 が最小となるときの調波構造モデル数を推定同時発話者数とする．

(注2)：2.2 節では，いずれも $g(x \mid \mu_k+\log n,\sigma^2)$ を用いて計算される
(注3)：この場合，矩形的に分割された領域内で一様な分布となる．

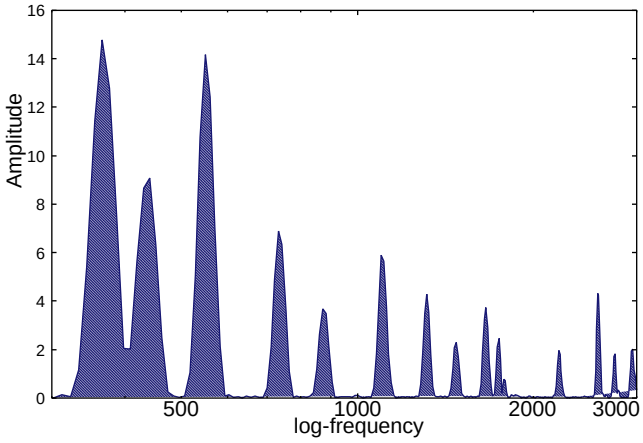

図 1　F_0 が 371Hz, 441Hz, 556Hz の 3 音による多重音スペクトル

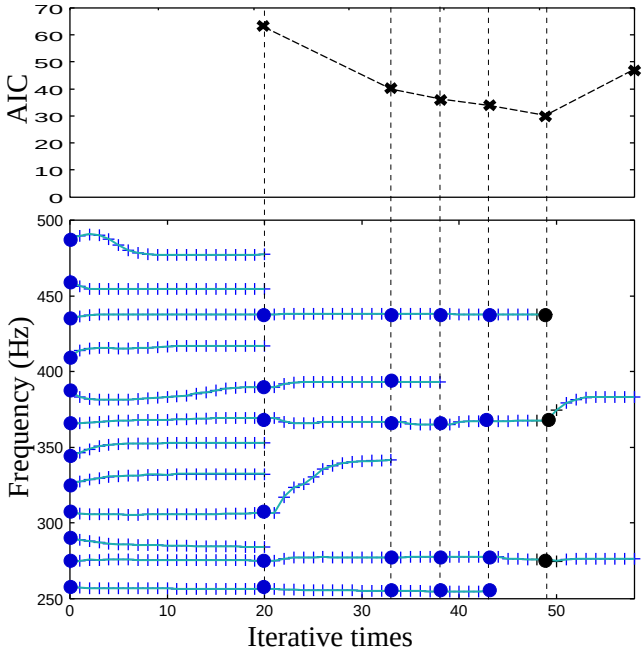

図 2　調波構造モデル数および基本平均の更新

（4） 全体のモデルが与える期待対数尤度に及ぼす重要度が低いと見なせる調波構造モデルを削減する．以下を行い，残った調波構造モデル数を $\bar{K}$ とする．$K = \bar{K}$ として (2) に戻る．

- ■ すべての調波構造モデルの中で重み w^k が最小のもを消滅させる．
- ■ 隣接する 2 つの基本平均がある一定閾値より近接した場合，w_k が小さい方を消滅させる．これは，1 つの極値に 2 つの基本平均が収束していると考えられるためである．

このプロセスを図 1 のスペクトルに対して実際に行った例を図 2 に示す．図 2 の下図における ‘+’ は基本平均の反復計算ごとの更新値，破線が (2) において収束判定により $\boldsymbol{\mu}$ と $\boldsymbol{w}$ が収束したと見なされた時点を表す．上図の折れ線グラフが各時点での AIC の値を表す．調波構造モデル数が 3 のときに AIC は最小値をとるため，この場合推定同時発音数は 3 となる．

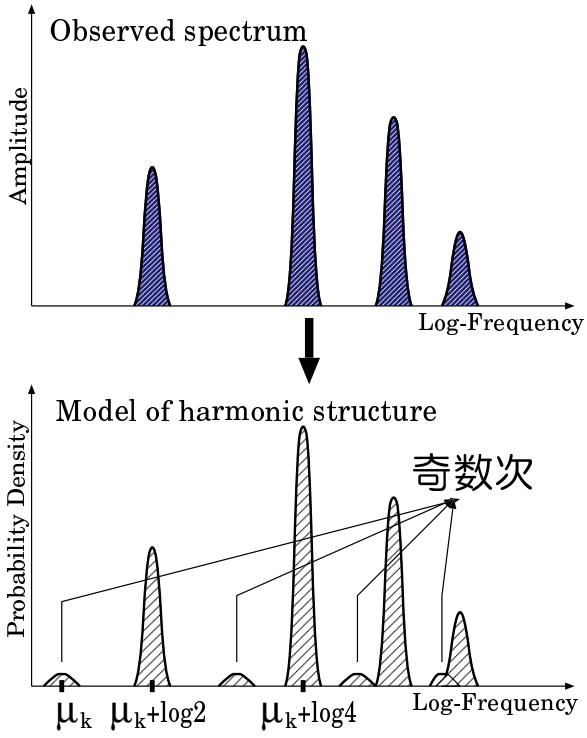

図 3　μ_k が真の F_0 の 1/2 であった場合の調波構造モデル

3.3　F_0 およびスペクトル形状検出プロセス

前述したプロセスにおいては，基本平均 μ_k が真の F_0 とその整数倍あるいは整数分の 1 倍の値のときも平均対数尤度を極大にすると考えられるため，得られる最尤推定量は必ずしも真の F_0 であるとは限らない．そこで，前述のプロセスで得られる基本平均 μ_k の最尤推定量が，真の F_0 あるいはその整数倍か整数分の 1 倍の値であることを前提とし，μ_k をそれぞれの値に順次置き換えながら何らかの規準に基づいて真の F_0 を検出する．

ここで，前節で与えた重み w_n^k に関する拘束を外し，すべての正規分布の重みに関して自由度を与えることにする(注4)．従って，重み w_n^k の最尤推定量は近似されたスペクトル形状，すなわち調波成分間の相対的な強度，を表すことになる．もし，置き換えた μ_k が真の F_0 より小さい場合，実際に存在する調波成分に対応する重みパラメータ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，過適応を起こしていると言える．例えば，μ_k が真の F_0 の 1/2 に対応した場合に重み w_n^k に関して最尤推定を行えば，偶数次の調波成分に比べて奇数次の調波成分が極端に小さい単一音のモデルとして表現されるはずである (図 3)．この観点から，前節のプロセス同様，AIC に基づいて真の F_0 を検出することができると考えられる．前節のプロセスにおいて残った調波構造モデルすべてについて以下の手順を行い，F_0 およびスペクトル形状の検出を行う．

（1） 調波構造モデル k における基本平均を $\mu_k + \log t$ に置き換える．ただし，t は初期値 1 の整数とする．このとき，上限がナイキスト周波数の対数である範囲内にとりうる正規分布の数を N_k^t とする．

(注4)：重みの総和は 1 なので，厳密には，自由度は 1 つを除く残りすべての重みに対して与えられる．

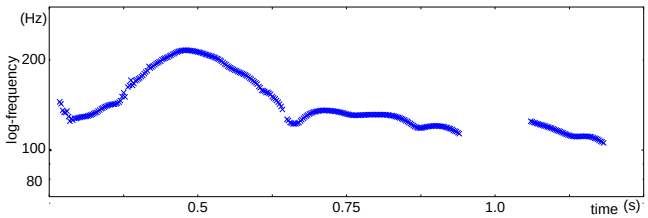

図 4　F_0 検出結果 1 (男性話者 1 人)

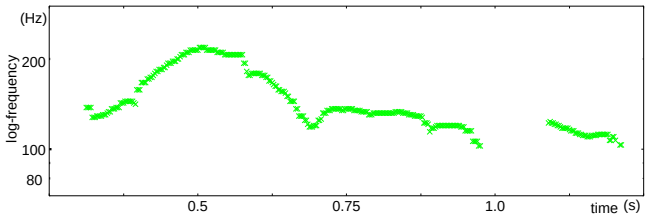

図 5　図 4 に対応する Reference F_0 パターン

表 1　話者 1 人の発話音声に対する推定正解率

Speech file	Accuracy(%)	
	Cepstrum	Proposed
'myisda01'	88.2	98.0
'myisda02'	88.4	99.0
'myisda03'	84.8	98.1
'myisda04'	85.1	92.4
'myisda05'	76.8	93.7
'fymsda01'	86.3	98.5
'fymsda02'	87.1	97.5
'fymsda03'	83.3	95.8
'fymsda04'	86.7	96.8
'fymsda05'	85.2	96.0

（2） EM アルゴリズムにより最尤パラメータを求める．ここでは，更新すべきパラメータは各正規分布の重み w_n^k だけである．M-ステップにおける w_n^k の更新値は以下で与えられる．

$$\bar{w}_n^k = \int_{-\infty}^{\infty} p_n^k(x)\, dx \qquad (16)$$

（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．t を 1 増やし，(1) に戻る．AIC が最小となるときの $\mu_k + \log t$ を推定 F_0 とする．

4. 動 作 実 験

上述した F_0 検出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 法と比較実験を行った．ATR 音声データベースより音声データ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．すべての音声信号はサンプリング周波数 12kHz でディジタル化され，フレーム長 64ms，フレームシフト 10ms のもとで Hamming 窓をかけて周波数解析 (FFT) を行い，スペクトル系列を得た．同時発話者数検出プロセスにおいて，初期調波構造モデル数は 4 とし，基本平均を配置する周波数範囲は 70Hz から 140Hz とした．また，すべての正規分布の分散の値は 0.45 とした．'fym' および 'myi' から始まる音声ファイル名はそれぞれ女性話者と男性話者による音声信号データをさす．評価基準として，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 と見なした．

表 2　話者 2 人の発話音声に対する推定正解率 (Cepstrum)

Speech files		Accuracy(%)	
File 1	File 2	Speaker 1	Speaker 2
'myisda01'	'myisda03'	63.7	63.1
'myisda01'	'myisda04'	45.7	51.6
'myisda02'	'myisda03'	63.3	50.1
'myisda02'	'myisda04'	59.4	42.1
'fymsda01'	'fymsda02'	57.7	54.0
'fymsda01'	'fymsda04'	53.1	41.0
'fymsda02'	'fymsda03'	52.9	59.6
'fymsda02'	'fymsda04'	64.9	64.7
'myisda01'	'fymsda03'	45.7	43.0
'myisda02'	'fymsda05'	55.0	44.5
'myisda03'	'fymsda04'	41.4	59.9
'myisda04'	'fymsda02'	64.9	50.6
'myisda05'	'fymsda03'	59.4	62.8
'myisda04'	'fymsda01'	62.0	71.7

表 3　話者 2 人の発話音声に対する推定正解率 (Proposed)

Speech files		Accuracy(%)	
File 1	File 2	Speaker 1	Speaker 2
'myisda01'	'myisda03'	90.1	83.0
'myisda01'	'myisda04'	92.8	81.3
'myisda02'	'myisda03'	88.2	85.7
'myisda02'	'myisda04'	84.4	87.6
'fymsda01'	'fymsda02'	90.7	84.3
'fymsda01'	'fymsda04'	85.3	82.6
'fymsda02'	'fymsda03'	79.2	90.3
'fymsda02'	'fymsda04'	86.2	92.6
'myisda01'	'fymsda03'	76.1	84.9
'myisda02'	'fymsda05'	74.8	92.8
'myisda03'	'fymsda04'	72.6	88.4
'myisda04'	'fymsda02'	86.3	85.5
'myisda05'	'fymsda03'	78.0	86.6
'myisda04'	'fymsda01'	79.0	86.6

4.1 話者一人による音声信号に対する実験結果

提案手法が多重 F_0 についてだけではなく単一 F_0 についても高い性能で推定できることを確認するため，話者一人による単一チャネル音声信号に対して動作実験を行い，単一 F_0 推定手法として広く知られる Cepstrum 法と推定正解率の比較を行った．推定正解率 (Accuracy) は，総フレーム数に対する gross error 以外のフレーム数の割合とした．

女性話者および男性話者それぞれの音声データにおける実験結果を Cepstrum 法の結果と併せて表 1 に示す．また，F_0 検出結果の例を図 4 に示し，対応する reference F_0 パターンを図 5 に示す．結果より，推定正解率 92.4%〜99.0%を得た．また，

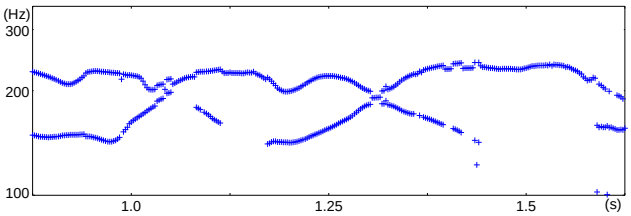

図 6　F_0 検出結果 1 (女性話者 2 人)

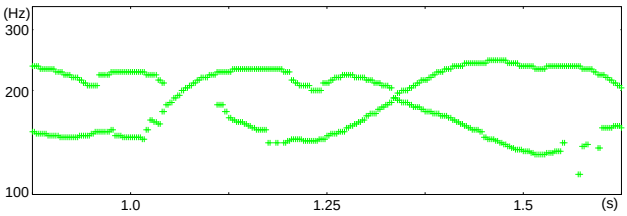

図 7　図 6 に対応する Reference F_0 パターン

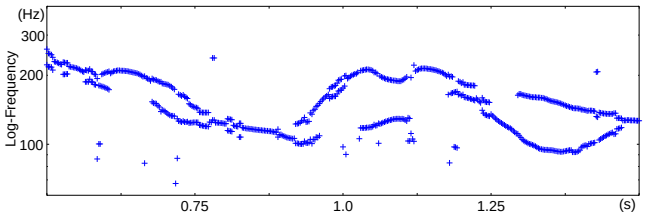

図 8　F_0 検出結果 2 (男性話者 2 人)

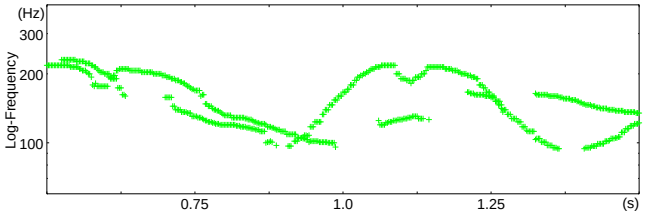

図 9　図 8 に対応する Reference F_0 パターン

すべての音声信号に対して Cepstrum 法に比べて推定正解率が高く，多重 F_0 だけではなく単一 F_0 推定手法としても高い性能であることが確認できた．

4.2　話者二人による同時発話音声信号に対する実験結果

次に，話者二人による単一チャネル音声信号に対して動作実験を行い，同様に Cepstrum 法との比較を行った．Cepstrum 法は複数話者による発話には原理的には適用できないため，厳密には客観評価の比較対象とはならないが，提案手法の客観評価のための参考基準としては十分であると考えた．

2 つの音声データの信号波形を人工的に加算したものを同時発話音声データとし，SSR(signal-to-signal ratio) は 0dB とした．Cepstrum 法による F_0 検出は，低ケフレンシー領域と高ケフレンシー領域を閾値により区分し，高ケフレンシーにおける 2 つのローカルピークを抽出することで行った．推定正解率は，同時発話されていると見なされるフレームを referenceF_0 パターンから判断し，同時発話時のフレーム総数に対する gross error 以外のフレーム数の割合とした．

Cepstrum 法の推定正解率を表 2，提案手法の推定正解率を表 3 に示す．また，提案手法の F_0 検出結果の例を図 6，8 に示し，それぞれに対応する reference F_0 パターンを図 7，9 に示す．Cepstrum 法では，推定正解率が 41.0%〜71.7%程度であったのに対し，提案手法では，72.6%〜92.8%であった．同時発話者数を事前に与えなくても AIC により F_0 および話者数を高精度に推定することができ，情報量規準が多重 F_0 推定に十分有効であることが確認できた．

5.　おわりに

本報告では，拘束つき混合正規分布のパラメータの最尤推定と情報量規準の 1 つである AIC に基づいて，同時発話音声の F_0，同時発話者数，スペクトル形状を検出する手法を提案し，話者一人および二人による音声信号に対してアルゴリズムの動作実験により性能の確認を行った．ただし，評価実験では Cepstrum 法を比較対象としたが，客観評価のために他の手法との比較も今後は行う必要がある．

今回は各フレーム独立に処理を行ったが，F_0 の時間連続性を考慮することによりさらに精度を高くできる可能性があるため，拡張方法を検討する．また，本報告の実験では検証しなかったが，提案手法ではスペクトル形状も検出することができるので，正弦波加算合成などにより音声分離実験も行いたい．

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] S. Godsill and M. Davy, “Baysian Harmonic Models for Musical Pitch Estimation and Analysis,” *Proc. ICASSP2002*, Vol. 2, pp. 1769–1772, 2002.

[4] 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.

[5] T. Virtanen and A. Klapuri, “Separation of Harmonic Sounds Using Linear Models for the Overtone Series,” *Proc. ICASSP2002*, Vol. 2, pp. 1757–1760, 2002.

[6] 西一樹, 安部素嗣, 安藤繁, “聴覚情景分析のための多重ピッチ追跡と調波分離アルゴリズム,” 計測自動制御学会論文誌, Vol. 34, No. 6, pp. 483–490, 1998.

[7] 安部素嗣, 安藤繁, “共有 FM-AM の時間周波数統合に基づく聴覚情景分析 (II) —最適な時間軸統合とストリーム音の再合成—,” 電子情報通信学会論文誌, Vol. J83-D-II, No. 2, pp. 468–477, 2000.

[8] M. Karjalainen and T. Tolonen, “Multi-pitch and Periodicity Analysis Model for Sound Separation and Auditory Scene Analysis,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 127–140, 2001.

[9] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *J. of Royal StatisticalSociety Series B*, Vol. 39, pp. 1-38, 1977.

[10] 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.

[11] 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,” 日本音響学会 2003 年春季研究発表会講演論文集, 3-7-3, pp. 837–838, 2003.

[12] 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.