Vol.2009-MUS-81 No.11
2009/7/30

多重奏中の歌声の基本周波数と音素を同時に推定可能な新たなフレームワーク

藤原弘将[†1,†2] 後藤真孝[†1] 奥乃博[†2]

本稿では，歌声の基本周波数（F0）と音素を同時に推定可能な新たな手法について述べる．本手法は，F0 と音素以外の歌声の他の要素も同時に推定できるように設計されているため，混合音中の歌声を認識するための新たなフレームワークと考えることができる．本手法は，歌声とその他の伴奏音が混ざった状態を，歌声を分離するのではなく，そのままの形で統計的にモデル化する．また，信頼性の高い歌声のスペクトル包絡を推定するために，様々な F0 を持つ複数の音の調波構造を使用する．F0 と音素の同時推定を，ポピュラー音楽 6 歌手 10 曲で評価した結果，提案法により F0 推定の性能が平均 3.7 ポイント，音素推定の性能が平均 6.2 ポイント向上することを確認した．

A novel framework for concurrently estimating F0 and phonemes of singing voice in polyphonic music

Hiromasa Fujihara,[†1,†2] Masataka Goto[†1] and Hiroshi G. Okuno[†2]

A novel method is described that can be used to concurrently recognize the fundamental frequency (F0) and phoneme of a singing voice (vocal) in polyphonic music. This method can be considered as a new framework for recognizing a singing voice in polyphonic music because it is designed to concurrently recognize other elements of a singing voice, though this paper focuses on the F0 and voiced phoneme. Our method stochastically models a mixture of a singing voice and other instrumental sounds without segregating the singing voice. It can also estimate a reliable spectral envelope by estimating it from the harmonic structure of many voices with various F0s. The experimental results of F0 and phoneme recognition with 10 popular-music songs by 6 singers showed that our method improves the recognition accuracy by 3.7 points for F0 estimation and 6.2 points for the phoneme recognition.

1. はじめに

音楽は，産業的にも文化的にも重要なコンテンツであり，その中でも歌声は重要な役割を果たしている．本稿では，混合音中の歌声の歌詞（音素）と基本周波数（F0）を同時に認識するための手法，W-PST（Weighted composition of Probabilistic Spectral Template）法を提案し，F0 推定と音素認識の実験によりその有効性を確認する．本稿では歌詞と F0 についてのみ触れるが，提案する手法は声質（歌手名）など歌声のその他の要素の認識にも適用可能であり，混合音中の歌声を扱うための新たなフレームワークと位置づけることができる．

歌詞は歌い手が歌声によって伝えたい内容を表現し，F0 は楽曲の旋律を表すと同時に，歌手の技巧や表情なども表現するため，どちらも歌声を構成する重要な要素である．そのため，混合音中からこれらの要素を自動認識する技術は，音楽情報検索などにも応用可能で，重要な基礎技術となる．例えば，歌詞が認識できることで，歌詞が未知の楽曲を歌詞を手がかりに検索できる．また，音素の自動認識技術は，歌詞と音楽の時間的対応付けに適用でき，歌詞をカラオケのように表示する音楽プレイヤーや音楽ビデオのテロップ自動作成などに応用できる[1)]．歌声の F0 推定は，ボーカルパートの自動採譜やハミング検索などに応用可能である．さらに，ハミング検索に歌詞の情報を統合することで，ハミング検索の精度が向上することも報告されている[2)] など，歌詞と F0 を同時に推定することでさらに応用範囲が広まる．しかし，歌声は話し声に比べて，ビブラートや F0 の変化幅の広さ，歌手の感情表現などに起因する変動が多い上に，伴奏音が大音量で重畳するため，歌声（音素）の自動認識は非常に難しい研究課題である．

我々は，今までに音楽と歌詞の時間的対応付け手法[1),3)] と混合音中の歌声の F0 推定手法[4)] について研究してきた．これらの手法では共通して，混合音から調波構造を手がかりに音を分離し，それを統計的手法により識別するというアプローチをとっていた．具体的には，歌詞の時間的対応付けの場合，既存手法によって推定された歌声の F0 の音がどの音素であるかを識別し，歌声の F0 推定の場合，各時刻の周波数成分の候補が歌声であるかそれ以外の音であるかを識別していた．

しかし，それらの手法は下記の 2 つの問題点を抱えていた．

分離の問題 歌声の認識性能が，その前段に行われる分離の性能に大きく依存していた．そのため，F0 推定や，分離の際にスペクトルから調波成分を選択する処理の誤りが，性

†1 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
†2 京都大学
Kyoto University

Vol.2009-MUS-81 No.11
2009/7/30

能に悪影響を与えていた．また，歌声とノイズの S/N 比や歌声の歪み度合いなどの情報を含んでいる背景雑音（分離対象の音以外の音）を，分離の過程で捨ててしまっていた．

スペクトル包絡推定の問題　従来の我々の手法では，スペクトル包絡を分離後の歌声の調波構造から推定しスペクトル包絡同士の距離を計算することで，歌声を認識していた．しかし，調波構造の各倍音成分は元のスペクトル包絡から F0 の整数倍の周波数成分をサンプリングしたものと考えることができるため，与えられた調波構造から元のスペクトル包絡を一意に復元することは原理的に不可能であった．そのため，例えば F0 が高い音など，調波構造の各倍音成分の谷間の幅が広い場合など，距離を正確に計算することが困難であった．

本稿では，これらの問題点を解決する新しい手法を提案する．この手法は，歌声を分離したり，単一の調波構造からスペクトル包絡を推定したりせず，観測されたスペクトルを伴奏音が重畳したありのままの形を確率的にモデリングする．さらに，学習の過程では，複数の調波構造を用いることで，より正確にスペクトル包絡を推定する．

2. 関 連 研 究

混合音中の歌詞または音素の認識に関する関連研究として，5)–10) がある．いずれの研究も，歌声を分離しているか，もしくは，そもそも伴奏の影響を考慮していないかで，前節で述べた問題は解決されていなかった．Gruhne らの歌声の音素認識の研究[5] では，文献 3) の手法と同様の手法で歌声を分離した後に統計的手法で音素を識別していた．伴奏を含む歌声と歌詞の時間的対応付けに取り組んだ[6)–9)] 研究では，隠れマルコフモデル（HMM）に基づく音声認識の標準的な手法（もしくはそれを簡略化した手法）を基本に，対象言語の特徴や楽曲の構造などのその他の情報を統合させることで性能の向上を図っていた．Chen ら[6)] は，歌声区間の検出と音響モデルの適応により，HMM を用いた強制アラインメントを高精度化していた．Iskandar ら[7)] は，各音節の継続時間調をモデル化することで，HMM を用いた強制アラインメントの探索範囲に制約をかけていた．Wong ら[8)] は，広東語のポピュラー音楽を対象にし，音の高低で意味を区別する声調言語である広東語の性質を利用することで，歌声の F0 を手がかりに対応関係を推定していた．Kan ら[9)] の開発したシステム LyricAlly では，対応付けの手がかりとして，歌詞中の各音素の発声時間長を利用していた．Lee ら[10)] らは，歌詞の構造（A メロ，サビなどの情報）が予めラベル付けされていると仮定して，音響信号から自動推定した楽曲構造と対応づけることで歌詞の段落単位で対応付けをしていた．

混合音中の歌声に対する F0 推定の研究として，11)–13) がある[⋆1]が，本研究のように歌声のスペクトル包絡をモデル化し学習することで歌声の F0 を推定しているものはなかった．Li ら[11)] は，既存の多重ピッチ解析手法の結果から，自己相関に基づく方法を用いて高域で最も優勢なピークを選択することで歌声の F0 を選択していた．Ryynänen ら[12)] は，F0 の変化の仕方や強度の情報などの低レベルの音響特徴量と，高レベルの音楽的文脈の情報を組み合わせて，歌声の F0 を推定していた．Sutton ら[13)] らは，歌声の変化の仕方と高域での優勢さという 2 種類の基準を HMM で統合することで歌声の F0 を推定していた．

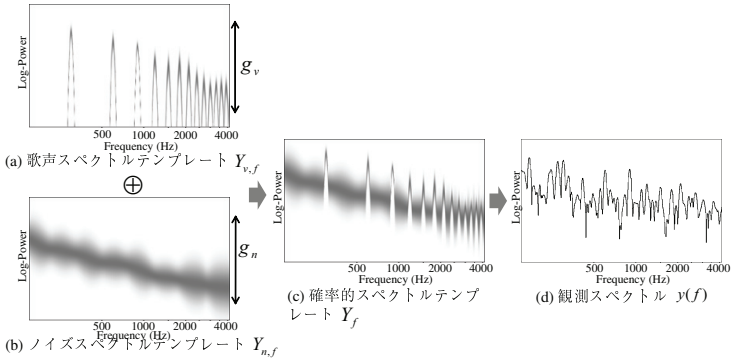


図 1　観測スペクトルの生成過程．図の濃淡は確率密度を表現する．重みパラメータ g_v と g_n を調整することで，様々な S/N 比のスペクトルを表現できる．

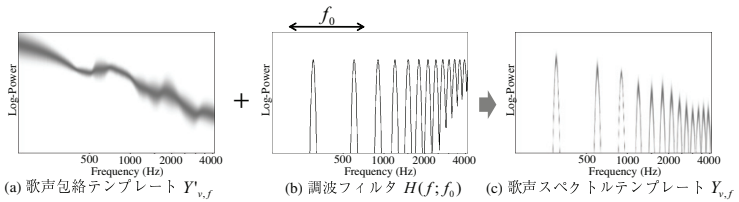


図 2　歌声スペクトルテンプレートの例．歌声包絡テンプレートと調波フィルタから生成される．

3. 歌声を認識するための新たなフレームワーク

図 1 (c) と (d) で示されるように，歌声を含む混合音のスペクトルがある確率分布の集合から生成されると仮定する．本稿では，それを**確率的スペクトルテンプレート**（Probabilistic

⋆1 歌声に限定しない一般のメロディに対する F0 推定の研究は他にもあるが，ここでは歌声に特化したもののみを紹介する．

Vol.2009-MUS-81 No.11
2009/7/30

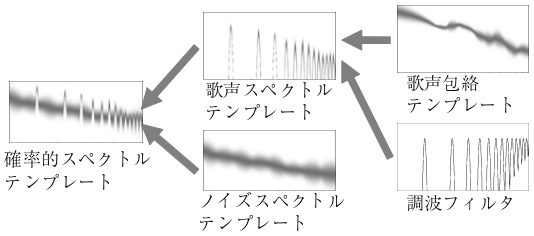


図 3 確率的スペクトルテンプレートの生成過程と名称のまとめ.

Spectral Template）と呼ぶ．ここで，スペクトルの各ビンのパワーはある確率分布に従い，その確率分布はスペクトルのビンごとに異なると考える．スペクトルの加法性を仮定すると，確率的スペクトルテンプレートは，歌声を表現するスペクトルテンプレート（図 1 (a)）と歌声以外の音を表現するスペクトルテンプレート（図 1 (b)）の線形軸上での加算で表現することができる．前者を**歌声スペクトルテンプレート**（Vocal Spectral Template），後者を**ノイズスペクトルテンプレート**（Noise Spectral Template）と呼ぶ．それらの 2 つのスペクトルテンプレートの加算の際に重みパラメータを導入し，重み付きで加算することで，様々な S/N 比のスペクトルを表現できる．さらに，ソースフィルターモデルを仮定すると，歌声スペクトルテンプレートは，スペクトル包絡を表現する**歌声包絡テンプレート**（Vocal Envelope Template）（**図 2** (a)）と駆動源の調波構造を表現する**調波フィルタ**（Harmonic Filter）（図 2 (b)）の積によって生成されると考えられる．調波フィルタの形状は，F0 の値をパラメータとして，コントロールできる．確率的スペクトルテンプレートの生成過程と名称のまとめを**図 3** に示す．ここで，この確率モデルのパラメータである調波フィルタの F0 と，歌声・ノイズスペクトルテンプレートのそれぞれの重みが定まれば，観測スペクトルのモデルに対する尤度を計算することができる．このモデルを用いると，各音素を表現する歌声包絡テンプレートをあらかじめ学習しておき観測スペクトルに対して最尤な歌声包絡テンプレートを選択することで音素認識ができ（**図 4**），最尤な F0 の値を推定することで F0 推定ができる．

本手法には，下記のような新規性がある．

- 本手法は，歌声を分離せず，ノイズ（伴奏音）が混在した状態をそのまま表現する．人間は歌声を分離せずにそのまま認識できることを考えると，人間の知覚の観点からも自然なやり方である．
- 本手法では，歌声と伴奏音の S/N 比を各フレームごとに推定可能なため，伴奏音の変動に対して頑健である．さらに，複数のノイズスペクトルテンプレートを用意し，最尤なものを選択することで，より頑健にすることができる．

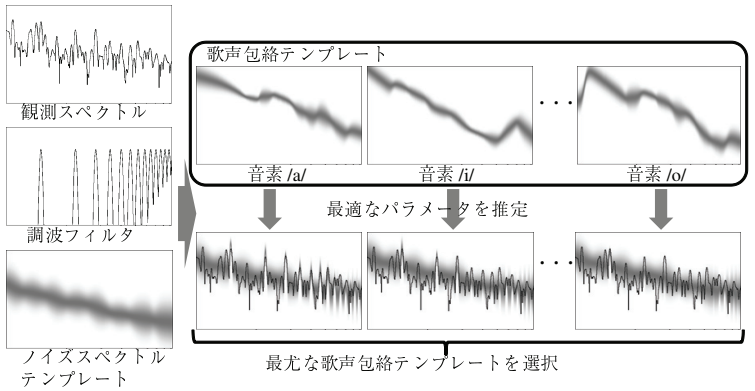


図 4 音素認識方法の概要.

- 本手法は，単一の調波構造からスペクトル包絡を推定しないため，高い F0 を持つ音に対しても頑健である．
- 本手法は，F0 を持たない無声子音など，他の音や音源に対しても，調波フィルタを用いない歌声スペクトルテンプレートを用意することで容易に拡張できる．

4. 定 式 化

本章では，3 節で述べた手法の具体的な定式化について述べる．本手法を実装するに当たって，下記の 3 つの手法を開発する必要がある．

(1) 確率的スペクトルテンプレートの表現方法.
(2) 2 つのスペクトルテンプレートの加算の計算方法.
(3) パラメータである，F0 とゲインを最適化する方法.

上記の問題に対して，本研究では下記のようなアプローチを取る．

(1) 確率的スペクトルテンプレートの各周波数ビンの分布として，対数正規分布を用いる．
(2) 対数正規分布に従う確率変数を加算した確率変数が，対数正規分布に従うと仮定する[⋆1].
(3) 準ニュートン法によりパラメータを最適化する．

4.1 確率的スペクトルテンプレート

歌声を含む混合音のスペクトル $y(f)$ は，確率変数 Y_f から生成されると仮定する．ただ

⋆1 一般には，対数正規分布に従う確率変数を加算した確率変数は対数正規分布に従わない.

し，f は対数軸での周波数を表し，s は対数軸でのスペクトルのパワーを表す．この確率変数（の集合）Y_f を確率的スペクトルテンプレートと呼ぶ．

次に，Y_f は次式により 2 つの異なるスペクトルテンプレートに分割できると仮定する．

$$Y_f = \log(\exp(Y_{v,f} + g_v) + \exp(Y_{n,f} + g_n)) \tag{1}$$

ただし，$Y_{v,f}$ は歌声のスペクトルを表し，歌声スペクトルテンプレートと呼ばれ，$Y_{n,f}$ は歌声以外の音（伴奏音）のスペクトルを表し，ノイズスペクトルテンプレートと呼ばれる．g_v と g_n はそれぞれのテンプレートの重みであり，それらを変化させることで歌声とその他の音の S/N 比を変化させることができる．なお，式 (1) においては，線形軸上でスペクトルの加法性を仮定している．

$Y_{v,f}$ と $Y_{n,f}$ が，次式のように，（対数周波数軸上で）正規分布に従うと仮定する．

$$Y_{v,f} \sim \mathcal{N}(y; \mu_{v,f}, \sigma^2_{v,f}) \tag{2}$$

$$Y_{n,f} \sim \mathcal{N}(y; \mu_{n,f}, \sigma^2_{n,f}) \tag{3}$$

ここで，$\mathcal{N}(y; \mu, \sigma^2)$ は，平均 μ，分散 σ^2 の正規分布である．さらに，ソースフィルターモデルを仮定することで，調波構造を持つ歌声 $Y_{v,f}$ は，次式のように，包絡の確率モデルと調波構造を表現するフィルタの対数軸上の加算で表現できると仮定する（図 2）．

$$Y_{v,f} = Y'_{v,f} + \log H(f; f_0) \tag{4}$$

$$\sim \mathcal{N}(y; \mu'_{v,f} + \log H(f; f_0), \sigma^2_{v,f}) \tag{5}$$

$$H(f; f_0) = \sum_h \mathcal{N}(f; \log f_0 + \log h, \sigma^2_H) \tag{6}$$

ここで，$Y'_{v,f} \sim \mathcal{N}(y; \mu'_{v,f}; \sigma^2_{v,f})$ は歌声のスペクトル包絡を表現する確率変数であり，歌声包絡テンプレートと呼ぶ．また，$H(f; f_0)$ は F0 の値が f_0 のフィルターを表現し，調波フィルタと呼ぶ．なお，調波フィルタ $H(f; f_0)$ は確率変数ではないことに注意が必要である．

以上をまとめると，歌声と伴奏音が混ざったスペクトルを表現する確率的スペクトルテンプレートは下記のように表される．

$$Y_f = \log(\exp(Y'_{v,f} + \log H(f; f_0) + g_v) + \exp(Y_{n,f} + g_n)) \tag{7}$$

$$\sim p_f(y; \theta_v, \theta_n, f_0, g_v, g_n) \tag{8}$$

$$\theta_v = (\mu_{v,f}, \sigma^2_{v,f}) \tag{9}$$

$$\theta_n = (\mu_{n,f}, \sigma^2_{n,f}) \tag{10}$$

4.2 スペクトルテンプレートの加算の近似

式 (1) で表される確率的スペクトルテンプレート Y_f は，解析的に計算することは困難であるので，正規分布を用いて近似計算する．関数 $l(x_1, x_2)$

$$l(x_1, x_2) = \log(\exp(x_1) + \exp(x_2)) \tag{11}$$

の $(x_1, x_2) = (\mu_{v,f} + g_v, \mu_{n,f} + g_n)$ における 2 次のテーラー展開は

$$l(x_1, x_2) \approx \frac{\exp(\mu_{v,f} + g_v)}{\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n)} x_1 + \frac{\exp(\mu_{n,f} + g_n)}{\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n)} x_2 + C \tag{12}$$

$$\mu_{v,f} = \mu'_{v,f} + \log H(f; f_0) \tag{13}$$

のように計算される．ただし，C は x_1 と x_2 とは独立な定数である．ここで，パラメータ g_v，g_n，f_0 が固定された場合，式 (12) が x_1 と x_2 の重み付き加算であることに注意すると，

$$Y_f = l(Y_{v,f}, Y_{n,f}) = \log(\exp(Y_{v,f}) + \exp(Y_{n,f})) \tag{14}$$

は，

$$Y_f \sim \mathcal{N}(y; \mu_f, \sigma^2_f) \tag{15}$$

$$\mu_f = \log(\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n)) \tag{16}$$

$$\sigma^2_f = \frac{(\exp(\mu_{v,f} + g_v))^2 \sigma^2_{v,f} + (\exp(\mu_{n,f} + g_n))^2 \sigma^2_{n,f}}{(\exp(\mu_{v,f} + g_v) + \exp(\mu_{n,f} + g_n))^2} \tag{17}$$

のように表現される．

4.3 音素と F0 の推定

このモデルを使って音素と F0 を認識するためには，まず，それぞれの音素 i を表現する歌声包絡テンプレート θ^i_v とノイズスペクトルテンプレート θ_n を準備する必要がある．観測スペクトル $y(f)$ が与えられたとき，次式により $y(f)$ に含まれる音素 $\hat{i}$ と F0 $\hat{F}_0$ を推定することができる．

$$(\hat{i}, \hat{F}_0) = \operatorname*{argmax}_{i, f_0} \max_{g_v, g_n} \int_f p_f(y(f); \theta^i_v, \theta_n, f_0, g_v, g_n) df \tag{18}$$

$$= \operatorname*{argmax}_{i, f_0} \max_{g_v, g_n} \int_f \log \mathcal{N}(y(f); u_f, \sigma^2_f) df \tag{19}$$

ただし，u_f と σ^2_f は，それぞれ式 (16) と (17) で定義される．また，本稿の対象外ではあるが，歌手名推定ができるように拡張したい場合は，各歌手ごとに歌声包絡テンプレートを用意することで実現できる．

4.4 準ニュートン法によるパラメータ最適化

式 (19) を計算するためのパラメータ $\theta = (g_v, g_n, f_0)$ の最適化には，BFGS（Broyden-Fletcher-Goldfarb-Shanno）公式に基づく準ニュートン法を使用する．準ニュートン法は山登り法の一種であり，反復的にパラメータを更新する．本モデルにおいて，最小化すべき目的関数 $Q(\theta)$ は，

$$Q(\theta) = -\int_f \log \mathcal{N}(y(f); u_f, \sigma^2_f) df \tag{20}$$

で表される．ただし，$y(f)$ は観測スペクトルである．

ニュートン法では，目的関数を現在のパラメータの周りの二次のテイラー展開で近似し，パラメータを逐次的に更新する．しかし，ニュートン法では，2 次のテイラー展開の計算に必要な 2 次の導関数のヘッセ行列が正定値であることを仮定しているが，この仮定は必ずしも成立しなかった．一方，準ニュートン法では，ヘッセ行列を直接計算せずに，パラメータの更新による 1 次の導関数の変化を用いて次式のように数値的に近似することで，安定した最適化が可能である．

$$\begin{aligned} B^{(k+1)} &= B^{(k)} \\ &+ \frac{(\nabla Q(\theta^{(k+1)}) - \nabla Q(\theta^{(k)}))(\nabla Q(\theta^{(k+1)}) - \nabla Q(\theta^{(k)}))^{\mathsf{T}}}{(\nabla Q(\theta^{(k+1)}) - \nabla Q(\theta^{(k)}))^{\mathsf{T}}(\theta^{(k+1)} - \theta^{(k)})} \\ &+ \frac{B^{(k)}(\theta^{(k+1)} - \theta^{(k)})(\theta^{(k+1)} - \theta^{(k)})^{\mathsf{T}}B^{(k)}}{(\theta^{(k+1)} - \theta^{(k)})^{\mathsf{T}}B^{(k)}(\theta^{(k+1)} - \theta^{(k)})} \end{aligned} \tag{21}$$

ただし，k は反復回数を表す．

パラメータは下記のように最適化できる．

Step 0 $k = 0$ と $B^{(0)} = I$ を設定し，$\theta^{(0)}$ を初期化する．

Step 1 $\theta^{(k+1)}$ を次式により更新する．

$$\theta^{(k+1)} = \theta^{(k)} - \alpha^{(k)}(B^{(k)})^{-1}\nabla Q(\theta^{(k)}) \tag{22}$$

$\alpha^{(k)}$ の値は，線形探索により決定する．

Step 2 式 (21) により $B^{(k+1)}$ を更新する．

Step 3 1 に戻る

5. 歌声包絡テンプレートの推定

式 (4) 中の歌声包絡テンプレート $Y'_{v,f}$ とノイズスペクトルテンプレート $Y_{n,f}$ は，学習データから推定する．一般に，調波構造を持つ歌声のスペクトルは，真のスペクトル包絡に対して，基本周波数の整数倍の周波数成分の点をサンプリングしたものと考えることができる．そのため，観測された歌声のスペクトル（調波構造）と，その元となるスペクトル包絡は一対多の関係になり得るので，単一フレームの調波構造から真のスペクトル包絡を推定することは困難である．本研究では，異なる F0 の値を持つ複数フレームの調波構造を用いることで，信頼性の高いスペクトル包絡を推定する．また，スペクトル包絡を一意に定めるのではなく，確率分布として推定するので，歌声の変動や学習データとテストデータの違いに対して頑健となる．

複数の調波構造からその元となるスペクトル包絡を推定する場合，フレームごとの音量の違いを考慮に入れる必要がある．そのため，本研究では各フレームの音量を正規化するためのパラメータを導入し，それも未知パラメータとして推定することでこの問題を解決する．

5.1 混合回帰分布

スペクトルテンプレートを表現するモデルとして，各回帰要素として線形回帰を使用した混合回帰モデル[14] を導入する．前章で述べたように，本手法においてはスペクトルテンプレートはある周波数 f における対数パワーの分布が正規分布で表現されるモデルを用いて定義される必要があるが，このモデルはその用件を満たしている．混合回帰モデルでは，スペクトルテンプレートの平均 $\mu_{v,f}$ と分散 $\sigma^2_{v,f}$ を

$$\mu_{v,f} = \sum_m G_m(f; \psi_m, \mu_m, \sigma^2_m)(a_m f + b_m) \tag{23}$$

$$\sigma^2_{v,f} = \sum_m G_m(f; \psi_m, \mu_m, \sigma^2_m)^2 \beta^2_m \tag{24}$$

として表現する．ただし，$G_m(f; \psi_m, \mu_m, \sigma^2_m)$ はゲート関数の出力で，次式で定義される正規化ガウス関数[15] を用いた．

$$G_m(f; \psi_m, \mu_m, \sigma^2_m) = \frac{\psi_m \mathcal{N}(f; \mu_m, \sigma^2_m)}{\sum_{m'} \psi_{m'} \mathcal{N}(f; \mu_{m'}, \sigma^2_{m'})} \tag{25}$$

このモデルにおいて，未知パラメータは $\{\psi_m, \mu_m, \sigma^2_m, a_m, b_m, \beta^2_m\}$ であり，EM（Expectation and Maximization）法により推定することが可能である．ただし，ψ_m は，$\psi_m \geq 0$ かつ $\sum_m \psi_m = 1$ である．

5.2 パラメータ推定

学習データとして与えられた I フレーム分の調波構造 $s_i (i = 1, \ldots, I)$ の h 次倍音の周波数 $f_{i,h}$ とその対数パワー $y_{i,h}$ が，

$$s_n = \{(f_{i,1}, y_{i,1}), \ldots (f_{i,h}, y_{i,h}), \ldots (f_{i,H_i}, y_{i,H_i})\} \tag{26}$$

として表されるとする．この時，最大化したい尤度関数は，次式で表される．

$$L = \sum_i^N \sum_h^{H_i} \mathcal{N}(y_{i,h} + k_i; \mu_{v,f_{i,h}}, \sigma^2_{v,f_{i,h}}) \tag{27}$$

ここで，k_i は各調波構造の音量を正規化するオフセットパラメータである．混合回帰モデルのパラメータと k_i を同時に最適化することは困難なので，それらを反復的に更新していく．

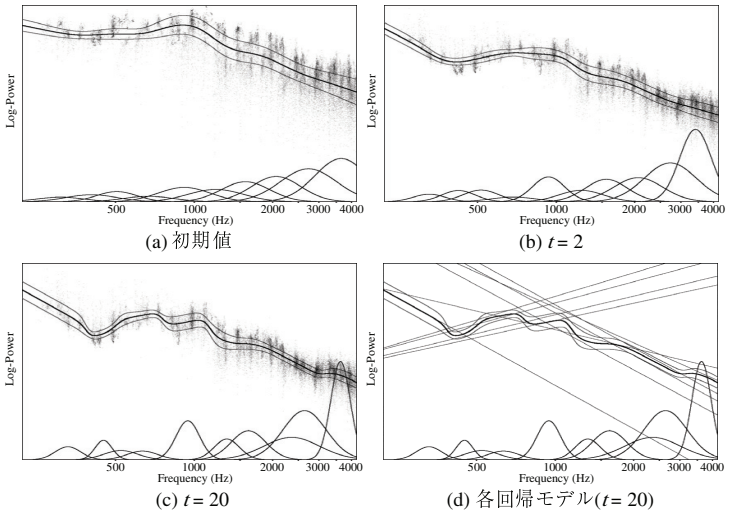
パラメータは下記の手続きで推定される．

Step 0 $k_i = 0$ とし，その他のパラメータの初期値を与える．

Step 1 混合回帰モデルのパラメータを EM 法により推定する．

Step 2 k_i を次式により更新する．

$$k_i = \frac{\sum_{h=1}^{H_i} \frac{\mu_{v,f_{i,h}} - y_{i,h}}{\sigma^2_{v,f_{i,h}}}}{\sum_{h=1}^{H_i} \frac{1}{\sigma^2_{v,f_{i,h}}}} \tag{28}$$



(a) 初期値　(b) $t = 2$　(c) $t = 20$　(d) 各回帰モデル($t = 20$)

図 5 混合回帰モデルのパラメータ推定の過程の一例．各図の中心の太い線は混合回帰モデルの平均を表し，その上下の細い 2 本の線は標準偏差を表す．背景の細かい点は学習データの調波成分を表し，各図の下部の複数の山は，ゲート関数 $G_m(f; \psi_m, \mu_m, \sigma_m^2)$ を表す．

Step 3　1 に戻る．

図 5 はパラメータの推定過程の例である．図より，更新を重ねることで学習データの各調波構造に対するオフセットパラメータ k_i が最適化されて，より分散の少ない回帰曲線が推定されていることが見てとれる．ノイズスペクトルテンプレートについては，$s_i(i = 1, \ldots, I)$ を調波構造でなくスペクトルそのものと考えることで，同様に推定できる．

6. 評 価 実 験

本章では，提案法の性能を確認するために行った評価実験について述べる．F0 と音素の同時推定の実験により提案法全体の性能を測り，F0 が与えられた条件下での音素推定の実験により音素推定単独の性能を評価した．

6.1 F0 と音素の同時推定

実験には，「RWC 研究用音楽データベース：ポピュラー音楽」[16] から選んだ 10 曲（男声 3 歌手，女声 3 歌手からなる）を用いた．音素推定の対象となる音素は日本語の 5 母音（/a/，/i/，/u/，/e/，/o/）とした．評価は，歌手ごとの 6 fold cross validation により行った．各楽曲に対して音素ラベルを手作業でアノテーションし，学習用音素ラベルと正解ラベルとして用いた．F0 についても同様に，手作業でアノテーションされた歌声の F0 データ[17] を正解ラベルとして用いた．音素，F0 共に，全体のフレーム数に対する正しく認識できたフレーム数の割合を正解率として評価した．ただし，対象の 5 母音を含むフレームのみで評価した．実験の際には，性別依存モデルを用いた．つまり，男声楽曲と女声楽曲で別々にテンプレートの集合（テンプレートモデル[⋆1]）を学習し，識別の際には，男声テンプレートモデルと女声テンプレートモデルの両方で尤度を計算し，尤度が高いテンプレートモデルの結果を採用した．

比較法として，F0 に関しては PreFEst[18] を，音素推定に関しては文献 3) の手法に基づいて分離した歌声から推定された MFCC を GMM により識別する手法を用いた．提案法及び比較法に関する分析条件を表 1 と 2 に示す．テンプレートの学習の際には，まず学習データとして使用する各楽曲に対して，歌声のみの音響信号と，歌声以外の伴奏音の音響信号（カラオケトラック）を準備した．次に，各楽曲から，各音素に対して一つの歌声包絡テンプレートと，一つのノイズスペクトルテンプレートを学習した．識別の際に，各音素に対して尤度を計算する際は，その音素に対応するすべての歌声包絡テンプレートと，すべてのノイズスペクトルテンプレートの組み合わせに対して尤度を計算し，最も尤度の高い組み合わせの尤度を採用した．F0 推定の比較法として採用した PreFEst は，各フレームの F0 の候補を計算する PreFEst-core と，それらの候補から時間的連続性を考慮して F0 を決定する PreFEst-backend からなるが，本稿では提案法において時間的連続性を考慮した処理を行っていないため，PreFEst-backend は用いず PreFEst-core のみで評価を行った．

実験結果を表 3 に示す．提案法により，10 曲の平均で音素推定は 6.2 ポイント，F0 推定は 3.7 ポイント性能が向上していることがわかる．音素推定では，10 曲中 7 曲で比較法より性能が向上している．特に No. 4 の楽曲では比較法では女声モデルの方が男声モデルより尤度が高くなっていたため，誤って女声モデルが使われてしまっているが，提案法では正しく男声モデルを選択できたので尤度が大幅に向上している[⋆2]．F0 推定に関しては，10 曲中 8 曲で比較法より性能が向上している．一方で，No. 9 の楽曲は，提案法で F0 推定の正解率が 22.2 ポイントと大幅に低下している．この楽曲では，伴奏に使われているギターが大音量で鳴っており，そのギターの F0 を誤って推定しまう場合が多かった．この F0 推定の誤りのために，音素認識においても比較法の方が性能が 4.1 ポイント高かった．この問題に対処するためには，ギターなどの歌声以外の音のテンプレートを準備し，それらのテンプレートに対する尤度と比較するなどのアプローチが有効であると考えられる．

⋆1 推定対象の複数の音素に対応する歌声包絡テンプレートと，ノイズスペクトルテンプレートの集合を，テンプレートモデルと呼ぶ．

⋆2 なお，比較法において性別非依存のモデルを使用した場合では，No. 4 以外の楽曲では性別依存モデルの場合より性能が低下し，10 曲の平均でも性別依存モデルより 1 ポイント低い正解率だった．

表 1 提案法の分析条件

スペクトル分析	サンプリング周波数	16 kHz
(連続ウェーブレット変換)	フレームシフト	10 msec
	周波数解像度	10 cent
	分析周波数帯域	60-4200 Hz
	マザーウェーブレット	ガボールウェーブレット
混合回帰モデル	混合数	10
調波フィルタ	σ_H^2	10 cent

表 2 比較法の分析条件

スペクトル分析	サンプリング周波数	16 kHz
(短時間フーリエ変換)	フレームシフト	10 msec
	フレームサイズ	25 msec
	窓関数	ハミング窓
MFCC	次元数	12
	メルフィルタバンクの次元数	24
GMM	混合数	32

表 3 音素と F0 の同時推定の実験結果 (正解率 [%])：提案法の結果における ↑ は，比較法より性能が向上した場合を表す．

			比較法		提案法	
楽曲 *	性別	歌手	音素認識	F0 推定	音素認識	F0 推定
No. 4	男	A	31.1**	62.6**	73.5↑	58.9
No. 11	男	A	56.5	65.6	57.6↑	71.5↑
No. 9	男	B	47.5	65.5	43.4	43.3
No. 12	男	B	62.8	76.8	63.9↑	77.6↑
No. 6	男	C	51.5	69.2	60.4↑	80.8↑
No. 2	女	D	69.5	71.6	68.5	86.3↑
No. 16	女	D	62.7	78.2	65.4↑	82.6↑
No. 7	女	E	60.0	73.8	67.2↑	82.7↑
No. 18	女	E	64.1	73.5	70.2↑	87.6↑
No. 14	女	F	44.1	79.1	42.3	82.0↑
平均			55.0	71.6	61.2↑	75.3↑

*RWC 研究用音楽データベース：ポピュラー音楽（RWC-MDB-P-2001）[16] の楽曲番号
** 異なる性別のモデルを誤って選択した楽曲

6.2 F0 が既知の条件下での音素推定

提案法の音素認識単体の性能を調べるため，F0 が既知の条件下での音素推定性能を評価した．下記の 3 通りの実験条件で評価を行った．

(i) 比較法 1 歌声の分離を行わず，伴奏が混在した状態のまま MFCC を抽出し GMM で識別した．

(ii) 比較法 2 F0 の正解を与え，前節の実験の比較法と同様に，文献 3) の手法で分離した歌声から MFCC を抽出し，GMM で識別した．

(iii) 提案法 F0 の正解を与え，本稿で提案した手法により音素を識別した．

条件 (ii) と (iii) は，F0 の正解を与えていることを除くと前章の実験と同様である．なお，F0 の正解とは，手作業でアノテーションされた歌声の F0 データ[17] を指す．

表 4 F0 が既知の条件下での音素推定の実験結果 (正解率 [%])：提案法の結果における ↑ は，比較法より性能が向上した場合を表す．

楽曲 *	性別	歌手	(i) 比較法 1	(ii) 比較法 2	(iii) 提案法
No. 4	男	A	31.1**	33.0**	64.3↑
No. 11	男	A	52.0	57.1	63.0↑
No. 9	男	B	30.0**	48.4	52.6↑
No. 12	男	B	33.8**	67.5	69.3↑
No. 6	男	C	42.6**	50.8	61.7↑
No. 2	女	D	59.1	70.7	70.7
No. 16	女	D	57.2	63.1	69.9↑
No. 7	女	E	54.4	62.3	70.2↑
No. 18	女	E	59.0	66.9	71.6↑
No. 14	女	F	40.4	43.9	46.2↑
平均			46.0	56.4	65.1↑

*RWC 研究用音楽データベース：ポピュラー音楽（RWC-MDB-P-2001）[16] の楽曲番号
** 異なる性別のモデルを誤って選択した楽曲

実験の結果を，表 4 に示す．提案法の精度は，比較法 1 と比べて 19.1 ポイント，比較法 2 と比べて 8.7 ポイント向上している．また，提案法により性能が低下している楽曲がないことがわかる．さらに，比較法ではいくつかの楽曲で誤った性別のモデルを選択しているが，提案法ではそのような楽曲がなかった．実験結果において，提案法（条件 iii）と比較法 2（条件 ii）で誤っていたフレームを比較したところ，提案法の不正解フレームの 52.6%は，比較法 2 では正しく識別されていることがわかった．これは，提案法と比較法を組み合わせることで，さらに性能が向上する可能性があることを示唆している．

7. ま と め

本稿では，多重奏の楽曲中の歌声の音素と F0 を同時に推定する手法について述べた．本手法の特徴は，歌声がその他の伴奏音と混ざった状態のスペクトルを，分離せずそのまま認識することにある．これは，人間は音を分離せずとも認識できるというアイデア[18] に基づいている．混合音を認識するための従来のやり方の多くは，構成するそれぞれの音を分離し，その後分離した音を認識するというアプローチだった．本研究のアプローチは背景のノ

イズに関する情報も活用するため，従来よりも性能を向上させることができる．

本手法は，音声認識の研究分野で知られる HMM 合成法[19] と共通点がある．それは，クリーン音声（歌声）のモデルとノイズのモデルを合成し，雑音下音声（歌声）のモデルを作成する点である．HMM 合成法では，合成は学習段階で行われるのであらかじめ用意しておいた S/N 比でしか合成できなかったが，提案法は各フレームで S/N 比の推定を行うのでノイズの変動に対してロバストになるという利点がある．

本研究の最終的な目標は，歌詞を自動的に認識するシステムを実現することである．今後は，その実現を目指して，本フレームワークを拡張していく予定である．例えば，本稿で扱った 5 母音のみでなく，無声子音も含めたすべての音素について有効性を確認していく予定である．また，本稿では，歌声が存在するという前提で音素と F0 の認識をしていたが，歌声が存在するかどうかを検出できるようにする必要がある．その他，現状では 1 フレームからなるテンプレートを，複数のフレームからなる 3 次元テンプレートに拡張することで，歌声の動的な特徴を表現することを考えている．

謝辞 本研究の一部は CrestMuse プロジェクト (JST CREST) の支援を受けた．

参 考 文 献

1) Fujihara, H. and Goto, M.: Three Techniques for Improving Automatic Synchronization between Music and Lyrics: Fricative Sound Detection, Filler Model, and Novel Feature Vectors for Vocal Activity Detection, *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2008)*, pp.69–72 (2008).
2) Suzuki, M., Hosoya, T., Ito, A., and Makino, S.: Music Information Retrieval from a Singing Voice Using Lyrics and Melody Information, *EURASIP Journal on Advances in Signal Processing*, Vol.2007 (2007).
3) Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T. and Okuno, H.G.: Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals, *Proc. ISM*, pp.257–264 (2006).
4) 藤原弘将，後藤真孝，奥乃　博：歌声の統計的モデル化とビタビ探索を用いた多重奏中のボーカルパートに対する音高推定手法，情報処理学会論文誌， Vol.49, No.10 (2008).
5) Gruhne, M., Schmidt, K. and Dittmar, C.: Phoneme recognition in popular music, *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pp.369–370 (2007).
6) Chen, K., Gao, S., Zhu, Y. and Sun, Q.: Popular Song and Lyrics Synchronization and Its Application to Music Information Retrieval, *Proceedings of the Thirteenth Annual Multimedia Networking and Computing (MMCN'06)* (2006).
7) Iskandar, D., Wang, Y., Kan, M.-Y. and Li, H.: Syllabic Level Automatic Synchronization of Music Signals and Text Lyrics, *Proceedings of the ACM Multimedia Conference*, pp.659–662 (2006).
8) Wong, C.H., Szeto, W.M. and Wong, K.H.: Automatic lyrics alignment for Cantonese popular music, *Multimedia Syst.*, Vol.4-5, No.12, pp.307–323 (2007).
9) Kan, M.-Y., Wang, Y., Iskandar, D., Nwe, T.L. and Shenoy, A.: LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals, *IEEE Trans. Audio, Speech, and Language Process.*, Vol.16, No.2, pp.338–349 (2008).
10) Lee, K. and Cremer, M.: Segmentation-based Lyrics-audio alignment using Dynamic Programming, *Proceedings of the 9rd International Conference on Music Information Retrieval (ISMIR 2002)*, pp.396–400 (2008).
11) Li, Y. and Wang, D.: Detecting pitch of singing voice in polyphonic audio, *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pp.III–17–20 (2005).
12) Ryynänen, M. and Klapuri, A.: Transcription of the Singing Melody in Polyphonic Music, *Proc. ISMIR 2006*, pp.222–227 (2006).
13) Sutton, C., Vincent, E., Plumbley, M.D. and Bello, J.P.: Transcription of vocal melodies using voice characteristics and algorithm fusion, *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX2006)* (2006).
14) Jacobs, R.J., Jordan, M., Nowlan, S.J. and Hinton, G.E.: Adaptive mixtures of local experts, *Neural Computation*, Vol.3, pp.79–87 (1991).
15) Xu, L., Jordan, M.I. and Hinton, G.E.: An alternative model for mixtures of experts, *Advances in Neural Information Processing Systems 7*, pp.633–640 (1994).
16) Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pp.287–288 (2002).
17) Goto, M.: AIST Annotation for the RWC Music Database, *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pp. 359–360 (2006).
18) Goto, M.: A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals, *Speech Communication*, Vol.43, No.4, pp.311–329 (2004).
19) Gales, M. J.F. and Yound, S.: An improved approach to the hidden Markov model decomposition of speech and noise, *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*, pp.835–838 (1997).