情報処理学会論文誌 Vol. 51 No. 10 1995–2006 (Oct. 2010)

多重奏中の歌声の基本周波数と有声音素の同時推定手法

藤原弘将[†1,†2] 後藤真孝[†1] 奥乃博[†2]

本論文では，歌声の基本周波数（F0）と母音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 と母音音素だけでなく，歌手名や性別など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 と母音音素の同時推定を，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 Method for Concurrently Estimating F0 and Vowel Phoneme 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 vowel 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 including singer's name and gender, though this paper focuses on the F0 and vowel 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．本論文では，混合音中の歌声の歌詞と基本周波数（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 推定と母音音素認識の実験によりその有効性を確認する．本論文の実験では母音音素と F0 についてのみ扱うが，W-PST 法は子音や，声質（歌手名や性別）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．本論文の以下の記述では，簡単のため，音素とは母音音素のことを指すこととする．

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．たとえば，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うえに，伴奏音が重畳されるため，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

†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

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 つの問題点をかかえていた．

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，性能に悪影響を与えていた．また，歌声とノイズの SIR（Signal-to-Interference Ratio）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どうし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，たとえば F0 が高い音や，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．

本論文では，これらの問題点を解決する新しい手法，W-PST 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で確率的にモデリングする．さらに，学習の過程で，F0 の異なる複数のフレームの調波構造を用いることで，単一フレームの調波構造から推定する場合に比べて，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，文献 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，前章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研究[6)–9)] では，隠れマルコフモデル（Hidden Markov Model; 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）をもとに，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，各音節の継続時間長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] は，歌詞の構造（A メロ，サビなどの情報）があらかじ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付けることで歌詞の段落単位で対応付けをしていた．

歌声に限定しない一般のメロディに対する F0 推定の研究は多数あるが，ここでは歌声に特化したもののみを紹介する．混合音中の歌声に対する F0 推定の研究として，文献 11)–15) があるが，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] 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．Durrieu ら[14),15)] は，歌声とその他の伴奏音が混ざった状態をモデル化し，それらを分離することで，混合音中の歌声の F0 推定と分離を実現した．歌声のスペクトルをスペクトル包絡と調波構造を表すスペクトルの積で表現するという点で本研究と共通点があるが，学習に基づかず観測信号のみから F0 推定・分離を目指しているため，本研究とはアプローチが異なっている．

3. W-PST 法による歌声の認識

3.1 手法の概要

本論文では，歌声と伴奏音を含む音響信号を入力として，その歌声に含まれる F0 と音素を出力する問題を扱う．本節では，それを実現するための W-PST 法の概要と，必要な仮定について述べる．以降，本論文ではスペクトルとして，入力された音声信号を連続ウェーブレット変換することで得られるパワースペクトルを用いる．

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．本論文では，それを**確率的スペクトルテンプレート**（Probabilistic Spectral Template）と呼ぶ．ここで，スペクトルの各ビン（離散的に計算された周波数成分）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．

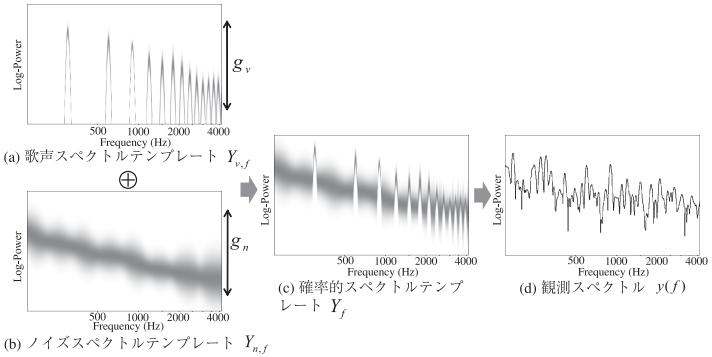


図 1　観測スペクトルの生成過程．図の濃淡は確率密度を表現する．重みパラメータ g_v と g_n を調整することで，様々な SIR のスペクトルを表現できる

Fig. 1　Generation process of the observed spectrum. The probability values are indicated by the darkness. Note that the SIR can be controlled by the gain parameters, g_v and g_n.

次に，パワースペクトルの加法性を仮定し，確率的スペクトルテンプレートを，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)）の加算で表現する．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．つまり，観測スペクトルを生成する音源を，歌声とそれ以外の音に分けて考え，それぞれが別々の確率的スペクトルテンプレートから独立に生成され，足しあわされることで観測スペクトルが生成されたと考える．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 SIR のスペクトルを表現できる．なお，W-PST（Weighted-composition of Probabilistic Spectral Template）法の名前の由来は，このようにスペクトルテンプレートを重み付きで合成する処理にある．

さらに，歌声スペクトルテンプレートは，**歌声包絡テンプレート**（Vocal Envelope Template）（**図 2** (a)）と**駆動音源関数**（Harmonic Excitation Function）（図 2 (b)）の積によって生成されると仮定する．この仮定は，ソースフィルタモデルを近似的に表現したものであり，歌声包絡テンプレートは音声の声道フィルタに，駆動音源関数は声帯振動のスペクトルに対応した概念である[⋆1]．駆動音源関数の F0 の値はパラメータとなっており，単一の歌

⋆1 ただし，あくまで近似的表現であり，ソースフィルタモデルを厳密に実現しているわけではない．

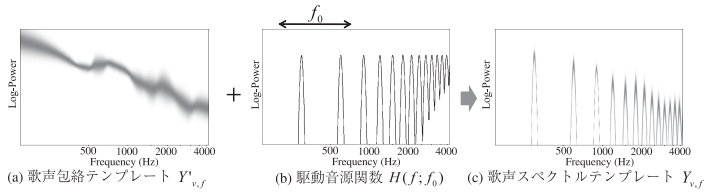


図 2　歌声スペクトルテンプレートの例．歌声包絡テンプレートと駆動音源関数から生成される

Fig. 2　Example of vocal spectral template, which generated from the vocal envelope template and the harmonic excitation function.

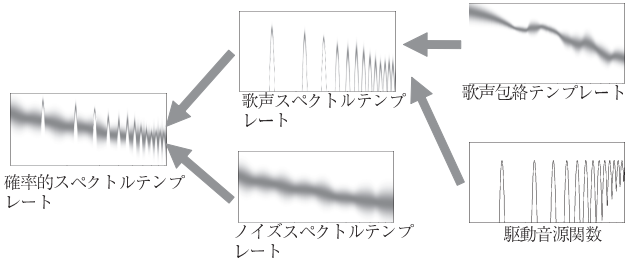


図 3　確率的スペクトルテンプレートの生成過程と名称のまとめ

Fig. 3　Summary of generation process of the probabilistic spectral template.

声包絡テンプレートから様々な F0 の歌声スペクトルテンプレートが生成できる．本論文では，ある観測スペクトルの元となった歌声包絡テンプレートをスペクトル包絡と呼び，ある有声音素の音は，必ずその音素に対応した（1 つまたは複数の）歌声包絡テンプレートから生成されると仮定する（つまり音素/i/用の歌声包絡テンプレートから，音素/a/の音が生成されることはない）．

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駆動音源関数の 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．推定に使用する歌声包絡テンプレートは，単一フレームの調波構造から学習するのではなく，複数フレームの異なる

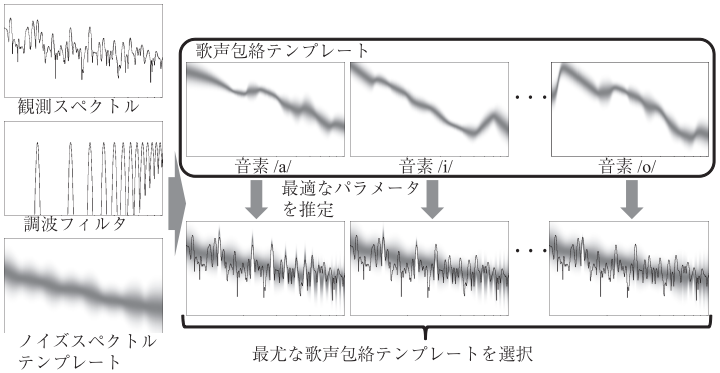


図 4　音素認識方法の概要
Fig. 4　Overview of phoneme estimation process.

F0 を持つ調波構造から前もって学習しておく．

3.2　手法の特徴

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．歌声の分離処理における誤りが認識処理に悪影響を及ぼさないという利点がある．
- 本手法では，歌声と伴奏音の SIR を各フレームごとに推定可能なため，伴奏音の音量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．
- 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という利点がある．従来手法は，調波構造の谷間など，歌声の周波数成分が観測されにくい周波数領域にも何らかのスペクトル包絡を推定して比較していたため，F0 が高い音やノイズに埋もれた音など，歌声の成分の有効な観測が少ない音から正確なスペクトル包絡を推定することが困難であった．本手法では，複数の調波構造を用いることで，より正確なスペクトル包絡を推定可能である．
- 本手法は，F0 と音素を同時に推定できるという利点がある．複数の音が混ざった信号から歌声の F0 を推定する場合に，テンプレートとして持つ歌声のスペクトル形状と近い音を選択するため，歌声以外の音源の音に惑わされずに歌声の F0 を推定できる．逆に，音素の推定の際に F0 を考慮に入れたモデル化を行うため，歌声の周波数成分が有効に観測される領域を使用でき，歌声以外の音の周波数成分の影響を受けにくくなる．
- 本手法は，母音音素以外の有声音に関しても適用可能である．また，F0 を持たない無声子音など，他の音や音源に対しても，駆動音源関数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 章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あたって，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とる．

(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近似する（なお，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）．
(3)　準ニュートン法によりパラメータを最適化する．

(1)，(2) のように，スペクトルを生成する確率分布を対数正規分布だと仮定・近似することで，観測スペクトルとテンプレートに対する尤度が，対数スペクトルの差の 2 乗誤差と等価なものになる．これは，音声認識で使用される距離尺度であるケプストラムのユークリッド距離とも等価となり，スペクトルの距離尺度としては自然なものとなる．

なお，本章ではまず識別対象のあらゆる有声音素を歌声として述べ，音素の種類によらない一般的な定式化を説明する．その後，4.3 節において，複数の異なる種類の音素を扱う方法について述べる．

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．ただし，f は対数軸での周波数を表し，$y(f)$ 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．

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

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 SIR を変化させることができる．なお，式 (1) においては，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．

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

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

ここで，$\mathcal{N}(\mu, \sigma^2)$ は，平均 μ，分散 σ^2 の正規分布を表す．

さらに，調波構造を持つ歌声を表現する確率変数 $Y_{v,f}$ 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スペクトルの加算で表現できると仮定する（図 2）．3 章で述べたように，これはソースフィルタモデルの近似的表現である．

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

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

$$H(f; f_0) = \log\left(\sum_h \exp(-(\log f_0 + \log h - \log f)^2 / 2\theta_H^2)\right) \tag{6}$$

ここで，$Y'_{v,f} \sim \mathcal{N}(\mu'_{v,f}, \sigma^2_{v,f})$ 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 の声帯振動のスペクトルを表現し，駆動音源関数と呼ぶ（図 2 (b)）．なお，駆動音源関数 $H(f; f_0)$ 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変数 Y_f は下記のように表される．

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

確率変数 Y_f はパラメータ（θ_v, θ_n, f_0, g_v, g_n）に依存する．ただし，θ_v と θ_n は，以下の組とし，それぞれ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パラメータを表す．

$$\theta_v = (\mu'_{v,f}, \sigma^2_{v,f}) \tag{8}$$

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

以降の説明では，便宜的に確率変数 Y_f が従う確率密度関数を $p_f(y; \theta_v, \theta_n, f_0, g_v, g_n)$ と記す．

4.2　スペクトルテンプレートの加算の近似

式 (7) で表される確率的スペクトルテンプレート Y_f の確率密度関数は，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 $l(x_1, x_2)$

$$l(x_1, x_2) = \log(\exp(x_1) + \exp(x_2)) \tag{10}$$

の $(x_1, x_2) = (\mu'_{v,f} + H(f; f_0) + g_v, \mu_{n,f} + g_n)$ における 1 次のテーラー展開は

$$l(x_1, x_2) \approx \frac{\exp(\mu'_{v,f} + H(f; f_0) + g_v)}{\exp(\mu'_{v,f} + H(f; f_0) + g_v) + \exp(\mu_{n,f} + g_n)} x_1 + \frac{\exp(\mu_{n,f} + g_n)}{\exp(\mu'_{v,f} + H(f; f_0) + g_v) + \exp(\mu_{n,f} + g_n)} x_2 + C \tag{11}$$

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 (11) が x_1 と x_2 の重み付き加算であることに注意すると，$Y_f = l(Y'_{v,f} + H(f; f_0) + g_v, Y_{n,f} + g_n)$ が従う確率密度関数 $p_f(y; \theta_v, \theta_n, f_0, g_v, g_n)$ は，

$$p_f(y; \theta_v, \theta_n, f_0, g_v, g_n) \approx \mathcal{N}(y; \mu_f(\theta_v, \theta_n, f_0, g_v, g_n), \sigma^2_f(\theta_v, \theta_n, f_0, g_v, g_n)) \tag{12}$$

$$\mu_f(\theta_v, \theta_n, f_0, g_v, g_n) = \log(\exp(\mu'_{v,f} + H(f; f_0) + g_v) + \exp(\mu_{n,f} + g_n)) \tag{13}$$

$$\sigma^2_f(\theta_v, \theta_n, f_0, g_v, g_n) = \frac{(\exp(\mu'_{v,f} + H(f; f_0) + g_v))^2 \sigma^2_{v,f} + (\exp(\mu_{n,f} + g_n))^2 \sigma^2_{n,f}}{(\exp(\mu'_{v,f} + H(f; f_0) + g_v) + \exp(\mu_{n,f} + g_n))^2} \tag{14}$$

のように表現される．ただし，$\mathcal{N}(y; \mu, \sigma^2)$ は，平均 μ，分散 σ^2 の正規分布の確率密度関数を表す．ここで，2 つの確率変数 $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$，$Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ の重み付き和 $a_1 X_1 + a_2 X_2$（ただし，a_1，a_2 は定数）は，正規分布 $\mathcal{N}(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$ に従うことを用いた．これにより，対数軸上で正規分布に従うスペクトルテンプレートの重み付き加算を，正規分布で近似することができる．

本近似では，$\mu'_{v,f} + H(f; f_0) + g_v$ と $\mu_{n,f} + g_n$ の距離が近い場合に，近似の精度が低下する．しかし，実際は，歌声スペクトルテンプレートは調波構造の影響で周波数帯域ごとの上下の変動が激しいため，$\mu'_{v,f} + H(f; f_0) + g_v$ と $\mu_{n,f} + g_n$ は多くの f では離れた値をとる．そのため，スペクトル全体で尤度を計算する場合には，近似の精度が低い周波数帯域は多くなく，誤差の影響は小さくなる．

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 $\theta_{i,v}$ 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，$y(f)$ に含まれる音素と F0 の最尤推定値の組 $\{\hat{i}, \hat{f}_0\}$ を，次式により推定することができる．

$$\{\hat{i},\hat{f}_0\} = \underset{i,f_0}{\operatorname{argmax}} \max_{g_v,g_n} \sum_f \log p_f(y(f);\theta_{i,v},\theta_n,f_0,g_v,g_n) \tag{15}$$

$$\approx \underset{i,f_0}{\operatorname{argmax}} \max_{g_v,g_n} \sum_f \log \mathcal{N}(y(f);u_f(\theta_{i,v},\theta_n,f_0,g_v,g_n),\sigma_f^2(\theta_{i,v},\theta_n,f_0,g_v,g_n)) \tag{16}$$

ただし，$u_f(\theta_{i,v},\theta_n,f_0,g_v,g_n)$ と $\sigma_f^2(\theta_{i,v},\theta_n,f_0,g_v,g_n)$ は，それぞれ式 (13) と (14) で定義される．また，本論文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，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 (16) を計算するためのパラメータ $\lambda = (g_v, g_n, f_0)$ 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 $Q(\lambda)$ は，

$$Q(\lambda) = -\sum_f \log \mathcal{N}(y(f);u_f(\theta_{i,v},\theta_n,f_0,g_v,g_n),\sigma_f^2(\theta_{i,v},\theta_n,f_0,g_v,g_n)) \tag{17}$$

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 2 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．

$$\begin{aligned} B^{(k+1)} = B^{(k)} &+ \frac{(\nabla Q(\lambda^{(k+1)}) - \nabla Q(\theta^{(k)}))(\nabla Q(\lambda^{(k+1)}) - \nabla Q(\lambda^{(k)}))^{\mathsf{T}}}{(\nabla Q(\lambda^{(k+1)}) - \nabla Q(\lambda^{(k)}))^{\mathsf{T}}(\lambda^{(k+1)} - \lambda^{(k)})} \\ &+ \frac{B^{(k)}(\lambda^{(k+1)} - \lambda^{(k)})(\lambda^{(k+1)} - \lambda^{(k)})^{\mathsf{T}}B^{(k)}}{(\lambda^{(k+1)} - \lambda^{(k)})^{\mathsf{T}}B^{(k)}(\lambda^{(k+1)} - \lambda^{(k)})} \end{aligned} \tag{18}$$

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，$\lambda^{(0)}$ を初期化する．

Step 1　$\lambda^{(k+1)}$ を次式により更新する．

$$\theta^{(k+1)} = \lambda^{(k)} - \alpha^{(k)}(B^{(k)})^{-1}\nabla Q(\lambda^{(k)}) \tag{19}$$

$\alpha^{(k)}$ の値は，線形探索により決定する．

Step 2　式 (18) により $B^{(k+1)}$ を更新する．

Step 3　Step 1 に戻る

なお，実際には f_0 の最適化は，ローカルピークが多くあるため初期値依存性が高く，初期値近傍の値しか探索されない．そのため，まず櫛形フィルタによる手法などの簡便な手法により f_0 の値の候補をいくつか計算し，すべての候補を別々に Step 0 における初期値として使用し，複数の結果の中で最も尤度が高い結果を選択する．一方で，g_v と g_n の最適化は，ローカルピークが少なく容易である．これは，g_v と g_n は，歌声とノイズの SIR を表す項の g_r と全体のゲイン g_t に分割して，

$$g_v = g_t + g_r \tag{20}$$

$$g_n = g_t \tag{21}$$

のように扱うことができ，g_r と f_0 が決まった場合，g_t は解析的に計算することができることが理由である．g_v と g_n の初期値は，上記の式で $g_r = 0$ のとき（すなわち $g_v = g_n$）の最尤な g_t の値を解析的に計算し，その値に設定した．本論文における実験では，多くても 20 回程度の反復で収束することを確認した．

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 1 対多の関係になりうるので，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　混合回帰モデル

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[16] 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，このモデルはその要件を満たしている．混合回帰モデルは任意の非線形回帰を複数の線形回帰によって近似するモデルで，スペクトル包絡の形状について仮定をおかず，学習データのみに基づいてスペクトル包絡を推定する．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

$$\mu_{v,f} = \sum_{m=1}^{M} G_m(f;\psi_m,\mu_m,\sigma^2_m)(a_m f + b_m) \tag{22}$$

$$\sigma^2_{v,f} = \sum_{m=1}^{M} G_m(f;\psi_m,\mu_m,\sigma^2_m)^2\beta^2_m \tag{23}$$

として表現する．ただし，M は混合数を表す．また，$G_m(f;\psi_m,\mu_m,\sigma^2_m)$ 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[17] を用いた．

$$G_m(f;\psi_m,\mu_m,\sigma^2_m) = \frac{\psi_m\mathcal{N}(f;\mu_m,\sigma^2_m)}{\sum_{m'=1}^{M}\psi_{m'}\mathcal{N}(f;\mu_{m'},\sigma^2_{m'})} \tag{24}$$

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 $\sum_{m=1}^{M}\psi_m = 1$ である．

5.2　パラメータ推定

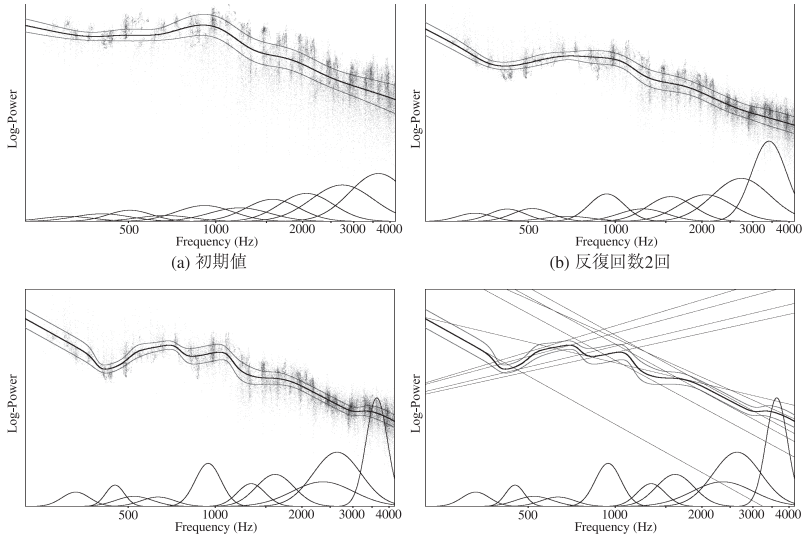
学習データとして与えられた I フレーム分の調波構造 s_i（$i = 1,\ldots,I$）の h 次倍音の周波数 $f_{i,h}$ とその対数パワー $y_{i,h}$ が，

$$s_i = \{(f_{i,1},y_{i,1}),\ldots,(f_{i,h},y_{i,h}),\ldots,(f_{i,H_i},y_{i,H_i})\} \tag{25}$$

として表されるとする．このとき，最大化したい尤度関数は，次式で表される．

$$L = \sum_{i}^{I}\sum_{h}^{H_i}\mathcal{N}(y_{i,h}+k_i;\mu_{v,f_{i,h}},\sigma^2_{v,f_{i,h}}) \tag{26}$$

ここで，k_i は各調波構造の音量をフレーム間で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．



(a) 初期値　(b) 反復回数2回　(c) 反復回数20回　(d) 各回帰モデル（反復回数20回）

図 5　混合回帰モデルのパラメータ推定の過程の一例．t は反復回数を表す．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，ゲート関数を構成する各ガウス分布とその重みの積 $\psi_m\mathcal{N}(f;\mu_m,\sigma^2_m)$ を表す

Fig. 5　Example of parameter estimation of the mixture of experts (MoE), where t represents the number of iteration. The middle line represents the mean of the MoE and the two thin lines around it represent the standard deviation. The background minute dots represent the harmonic components of the training data. A product of each Gaussian and its weight of the gating functions, $\psi_m\mathcal{N}(f;\mu_m,\sigma^2_m)$, are drawn at the bottom.

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に対して後述のように初期値を与える．

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．

$$k_i = \frac{\sum_{h=1}^{H_i}\dfrac{\mu_{v,f_{i,h}} - y_{i,h}}{\sigma^2_{v,f_{i,h}}}}{\sum_{h=1}^{H_i}\dfrac{1}{\sigma^2_{v,f_{i,h}}}} \tag{27}$$

Step 3　Step 1 に戻る．

k_i 以外のパラメータの初期値として，周波数軸の定義域（本論文の実験では 60 Hz〜4,200 Hz）を M 等分し，m 番目の分割について，$(f_{i,h}, y_{i,h})$ の回帰係数を計算したものを a_m と b_m の初期値に，$f_{i,h}$ の平均と分散を μ_m と σ_m^2 の初期値に設定し，ψ_m の初期値は $\frac{1}{M}$ とした．本論文の実験では，多くても 30 回程度の反復で学習は収束することを確認した．

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」[18] 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[19] を正解ラベルとして用いた．音素，F0 ともに，識別対象の 5 母音が含まれるフレームのみを用い，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．音素については，正解ラベルと同じ音素を出力した場合に正解と判断し，F0 については，正解ラベルと出力した F0 の差が 50 cent（半音の半分の音程差）以下の場合に正解と判断した．実験に使用した 10 曲と，各楽曲の対象 5 母音を含むフレームの総数，正解ラベルの F0 の平均，標準偏差，最小値，最大値を**表 1** に示す．また，音素ごとの F0 の分布に大きな違いはなかった．

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）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．なお，これは歌手の性別を推定していることに相当する．ただし，テンプレートモデルとは，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指す．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，1 つの歌声包絡テンプレートと，1 つのノイズスペクトルテンプレートを学習した．たとえば，男声 3 曲，女声 3 曲を学習データとしてテンプレートを作成する場合は，男声テンプレートモデルに対して音素/a/から/o/の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がそれぞれ 3 種類ずつ，合計 18 個学習され，女声テンプレートモデルに対しても同数のテンプレートが学習されることになる．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合せに対して尤度を計算し，最も尤度の高い組合せの尤度を採用した．

表 1　実験に使用した楽曲とその正解 F0 の性質
Table 1　Songs used in the experiments and their characteristics of ground-truth F0s.

楽曲 *	性別	歌手	総フレーム数 **	平均（Hz）	標準偏差（Hz）	最小値（Hz）	最大値（Hz）
No.4	男	A	6,358	279.4	89.7	110.0	460.8
No.11	男	A	9,794	261.5	55.0	92.5	395.7
No.9	男	B	12,828	250.1	69.0	90.4	594.5
No.12	男	B	9,061	288.2	49.5	133.1	396.6
No.6	男	C	6,712	323.8	49.4	161.0	415.4
No.2	女	D	6,675	357.3	60.9	174.6	544.8
No.16	女	D	9,643	338.3	64.4	141.8	499.8
No.7	女	E	14,745	416.9	76.5	237.2	608.0
No.18	女	E	12,898	424.6	93.3	218.7	706.6
No.14	女	F	8,238	336.7	64.9	201.7	502.5

* RWC-MDB-P-2001 [18] の楽曲番号
** 対象 5 母音を含むフレームのみ

比較法として，F0 に関しては PreFEst [20] 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．提案法および比較法に関する分析条件を**表 2** と**表 3** に示す．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，本論文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 4** に示す．図中の ↑ は，有意水準 5% で 2 群の比率の差の検定を行い，有意に性能が向上した楽曲を示す．No.12 の楽曲の F0 推定の結果は，性能の向上が有意でなかっ

表 2　提案法の分析条件
Table 2　Experimental conditions for proposed method.

スペクトル分析 （連続ウェーブレット変換）	サンプリング周波数 フレームシフト 周波数解像度 分析周波数帯域 マザーウェーブレット	16 kHz 10 msec 10 cent 60–4,200 Hz ガボールウェーブレット
混合回帰モデル	混合数	10
駆動音源関数	σ_H^2	10 cent

表 3　比較法の分析条件
Table 3　Experimental conditions for baseline method.

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

表 4　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，比較法より有意に性能が向上した場合を表す
Table 4　Experimental results for concurrent estimation of phonemes and F0 (%). ↑ represents a song of which the accuracy is improved by our method.

			比較法		提案（W-PST）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-MDB-P-2001 [18] の楽曲番号
** 異なる性別のモデルを誤って選択した楽曲

た．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分かる．同性の楽曲の学習データが 3 曲の楽曲（No.6 および No.14）と，4 曲の楽曲（上記 2 曲以外）では，認識率の向上度合いに有意な差は見受けられなかった．

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⋆1．また，表 4 より，性能向上の度合いが楽曲によってばらつきがあることが分かる．関連して，6.2 節でも述べるように，提案法と比較法では正解するフレームの傾向が異なる場合が多く見受けられた．今後，提案法がどのような楽曲やフレームに対して有効であるかを調査していく必要がある．

F0 推定に関しては，10 曲中 7 曲で比較法より有意に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 を誤って推定して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．また，No.4 の楽曲も F0 推定精度が低下している．この曲も同様に誤って異なる楽器の音を推定してしまう箇所が見られた．本手法では，歌声とそれ以外の音で分けてモデル化し，伴奏音についてはスペクトルの全体形状を大まかに表現するものになっている．そのため，比較法で用いた PreFEst などの音を要素に分解して推定する手法と比較して，観測した伴奏音の性質とモデルの伴奏音の性質が異なる場合に性能が低下してしまうことがあるのだと考えられる．この問題に対処するためには，様々な種類の伴奏音で学習したノイズスペクトルテンプレートを用意しておくアプローチや，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．

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．

⋆1 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 5　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，比較法より有意に性能が向上した場合を表す

Table 5　Experimental results for phoneme estimation (%). ↑ represents a song of which the accuracy is improved by our method.

楽曲 *	性別	歌手	(i) 比較法 1	(ii) 比較法 2	(iii) 提案（W-PST）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-MDB-P-2001 [18] の楽曲番号

** 異なる性別のモデルを誤って選択した楽曲

(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（W-PST）法　F0 の正解を与え，本論文で提案した W-PST 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節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前述の歌声の F0 データ[19] を指す．

本実験の結果を，**表 5** に示す．前節の実験と同様に，図中の ↑ は有意水準 5%で 2 群の比率の差の検定を行い，有意に性能が向上した楽曲を示す．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分かる．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分かった．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

本論文では，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，W-PST 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，人間は音を分離しなくても認識できるというアイデア[20] 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．W-PST 法は，従来の音声認識技術における，特徴抽出と GMM による音響尤度計算の部分を置き換えることのできる手法であり，従来法と同様に HMM などを用いてフレーム間の連続性や言語制約を考慮した認識を行うことが可能である．

本手法は，音声認識の研究分野で知られる HMM 合成法[21] や Vector Taylor Seriers（VTS）法[22] 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 SIR でしか合成できなかったが，提案法は各フレームで SIR 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．VTS 法は，さらに対数正規分布の加算をテイラー展開より近似するという点でも本手法と共通点があるが，提案法は調波構造を持つ歌声をモデル化しているという点で異なっている．

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．たとえば，本論文で扱った 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，本論文では，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, *Proc. 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2008*), 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. IEEE International Symposium on Multimedia* (*ISM 2006*), 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, pp.3682–3693 (2008).
5) Gruhne, M., Schmidt, K. and Dittmar, C.: Phoneme recognition in popular music, *Proc. 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, *Proc. 13th 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, *Proc. 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, *Proc. 9th 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, *Proc. 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. 7th International Conference on Music Information Retrieval* (*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, *Proc. Music Information Retrieval Evaluation eXchange* (*MIREX 2006*) (2006).
14) Durrieu, J.-L., Richard, G. and David, B.: Singer Melody Extraction in Polyphonic Signals Using Source Separation Methods, *Proc. ICASSP 2008*, pp.169–172 (2008).
15) Durrieu, J.-L., Richard, G. and David, B.: An Iterative Approach to Monaural Musical Mixture De-Soloing, *Proc. ICASSP 2009*, pp.105–108 (2009).
16) 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).
17) Xu, L., Jordan, M.I. and Hinton, G.E.: An alternative model for mixtures of experts, *Advances in Neural Information Processing Systems*, Vol.7, pp.633–640 (1994).
18) 後藤真孝，橋口博樹，西村拓一，岡　隆一：RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol.45, No.3, pp.728–738 (2004).
19) Goto, M.: AIST Annotation for the RWC Music Database, *Proc. 7th International Conference on Music Information Retrieval* (*ISMIR 2006*), pp.359–360 (2006).
20) 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).
21) Gales, M.J.F. and Yound, S.: An improved approach to the hidden Markov model decomposition of speech and noise, *Proc. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 1997*), pp.835–838 (1997).
22) Moreno, P.J., Raj, B. and Stern, R.M.: A Vector Taylor Series Approach for Environment-Independent Speech Recognition, *Proc. ICASSP 1996*, pp.733–736 (1996).

(平成 21 年 8 月 6 日受付)
(平成 22 年 7 月 9 日採録)

藤原　弘将（正会員）

2005 年京都大学工学部情報学科卒業．2007 年同大学大学院情報学研究科知能情報学専攻修士課程修了．同年産業技術総合研究所に入所し，現在に至る．博士（情報学）．2010 年，京都大学大学院情報学研究科知能情報学専攻博士課程修了．音楽情報処理，音楽情報検索，音声情報処理に興味を持つ．平成 19 年度山下記念研究賞受賞．日本音響学会，電子情報通信学会各会員．

後藤　真孝（正会員）

1998 年早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．同年電子技術総合研究所に入所し，2001 年に改組された産業技術総合研究所において，現在，情報技術研究部門メディアインタラクション研究グループ長．統計数理研究所客員教授，筑波大学大学院准教授（連携大学院），IPA 未踏 IT 人材発掘・育成事業未踏ユースプロジェクトマネージャーを兼任．ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞，科学技術分野の文部科学大臣表彰若手科学者賞，情報処理学会長尾真記念特別賞等，25 件受賞．

奥乃　博（正会員）

1972 年東京大学教養学部基礎科学科卒業．日本電信電話公社，NTT，JST，東京理科大学を経て，2001 年より京都大学大学院情報学研究科知能情報学専攻教授．博士（工学）．この間，スタンフォード大学客員研究員，東京大学工学部客員助教授．人工知能，音環境理解，ロボット聴覚，音楽情報処理の研究に従事．1990 年度人工知能学会論文賞，IEA/AIE-2001，2005，2010 最優秀論文賞，IEEE/RSJ IROS-2001，2006 Best Paper Nomination Finalist，IROS-2008 Award for Entertainment Robots and Systems Nomination Finalist 2 件，第 2 回船井情報科学振興賞等受賞．本学会理事．人工知能学会，日本ロボット学会，日本ソフトウエア科学会，ACM，IEEE，AAAI，ASA 等各会員．