混合音中の歌声スペクトル包絡推定手法と歌声の声質変換への応用*

◎藤原弘将, 後藤真孝 (産総研)

1 はじめに

本稿では，混合音中の歌声の声質の変換手法について述べる．つまり，入力として伴奏を含む歌声の音響信号と変換先歌手の歌声の音響信号を取り，歌声の声質が変換された音響信号を出力する手法である．ここで声質とは，歌声のスペクトルの静的な形状のことを指し，基本周波数（F0）の動きなど，動的な成分は含まないものとする．

他の音が背景等として含まれないクリーンな話し声を対象とした声質変換は数多く研究がなされており [1, 2, 3]，これらの技術の一部は単独歌唱の歌声にも適用が可能である．また，河原らの開発した歌声分析合成システム STRAIGHT に基づく歌声のモーフィング [4] では，2 種類の単独歌唱の歌声をリアルタイムにモーフィングし，ある歌手の声質で別の歌手の歌い方の歌を作成することなどができる．また，この STRAIGHT のモーフィングを声質変換に応用した研究例もある [5]．しかし，これらの単独歌唱を対象とした技術は混合音には適用できず，伴奏を含む混合音中の歌声の声質変換は今まで扱われてこなかった．

混合音中の歌唱の声質変換を行う際の本質的難しさは，歌声を処理する際に伴奏音の影響を排除する必要があるだけではなく，歌声への処理が伴奏音に与える影響を排除する必要がある点である．そこで本研究は，藤原ら [6] によって提案された W- PST 法を応用して，歌声の周波数成分のみを操作することを可能にした．W-PST 法は，混合音中の歌声の F0 と音素を推定する手法で，伴奏音と歌声が混ざった状態としてモデル化し，歌声の周波数成分が優勢な帯域を同定することが可能である．しかし，W-PST 法では歌声のスペクトル包絡推定は，単独歌唱のデータを用いていた．声質変換の目的では，変換したい音響信号は伴奏が混在した混合音として与えたいことが多いので，そのままではスペクトル包絡推定ができない．そこで，本研究では，混合音の音響信号中の歌声のスペクトル包絡推定手法を新たに開発することで，この問題を解決した．

2 W-PST 法に基づく混合音中の歌声の声質変換

本手法は，変換元の音響信号と変換先の歌手の音響信号を入力とし，変換元の音響信号の歌声の声質を変換先の歌手のものに変換した音響信号を出力する．具体的には，W-PST 法 [6] により歌声とノイズの SIR（Signal-to-Interference Ratio）を推定することで，混合音のスペクトル中で歌声の周波数成分が優勢な周波数帯域を同定し，歌声の周波数成分のみを操作する．図 1 に声質変換処理の概要を示す．

変換先歌手の歌声の音響信号は，変換元のものと同じ楽曲で有る必要はなく，複数の楽曲でも良い．一方で，単独歌唱の音響信号である必要があり，また変換元の音響信号に含まれる歌声の母音が含まれている必要がある．変換元および変換先の音響信号について，音素と F0 のラベル（各時刻における音素名と F0 の値）が付与されていることを仮定し，変換先の音響信号は単独歌唱のものであると仮定する．ただし，音素と F0 のラベルは文献 [6] の手法で推定することが可能であり，自動推定したラベルを使用して声質変換を行うことに今後取り組む予定である．また，本章の以下の処理は母音区間に対してのみ行われる．実際は，子音にも個人性が存在するため，子音区間に対して処理を行うことは今後の課題である．

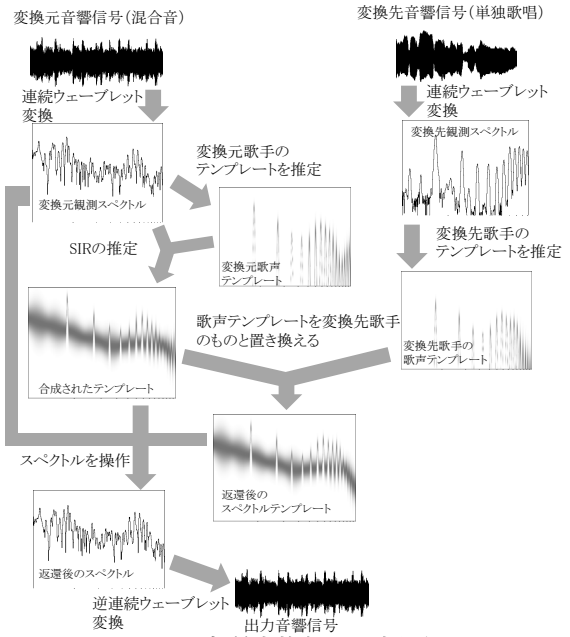


Fig. 1. 声質変換処理の概要.

2.1 ウェーブレット変換によるスペクトルの計算

まず，入力音響信号に対して連続ウェーブレット変換（CWT）をかけることで，スペクトログラムを計算する．本研究では，マザーウェーブレットとしてガボールウェーブレットを用いる．ガボールウェーブレットによる CWT は，入力音響信号を $x(t)$ とすると，下記のように定義される．

$$W(b,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) \overline{\Psi\left(\frac{t-b}{a}\right)} dt \quad (1)$$

$$\Psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{t^2}{2\sigma^2}\right) \exp\left(i\omega_0 t\right) \quad (2)$$

ただし，$\overline{\Psi(\cdot)}$ は，$\Psi(\cdot)$ の共役複素数を表す．b は時刻を表すパラメータで，$W(b,a)$ は全ての b について（つまり，離散信号の場合は全てのサンプルについて）計算される．a は周波数を表すパラメータで，$\frac{2\pi a}{w_0}$Hz に対応する．

次節以降で述べるテンプレートの推定処理では，計算時間を削減するために，10ms 間隔の離散的な b について（以降，フレームと呼ぶ）計算する．以降の処理は，それぞれの離散的な b の値について独立に行われるため，b の表記は省略し，対数パワースペクトル

*A spectral envelope estimation method for polyphonic music and its application to singing voice conversion. by FUJIHARA, Hiromasa and GOTO, Masataka (AIST)

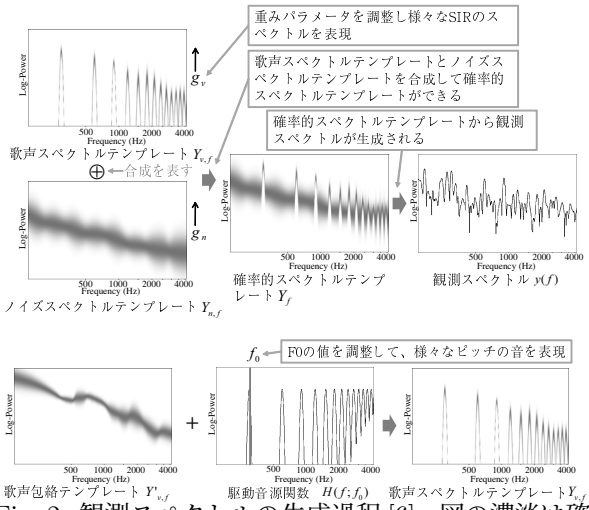


Fig. 2. 観測スペクトルの生成過程 [6]．図の濃淡は確率密度を表現する．

$y(f)$ を $y(f) = \log(|W(b,a)|)$ と表記する．ただし，a と f には $f = \log \frac{2\pi a}{w_0}$ という関係がある．

2.2 確率的スペクトルテンプレート

歌声を含む混合音の対数パワースペクトル $y(f)$ は，ある確率変数（の集合）Y_f から生成されると仮定する．この確率変数 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，2 つの異なるスペクトルテンプレート $Y'_{v,f}$ と $Y_{n,f}$ と，周波数に依存する関数 $H(f;f_0)$ を用いて，次式のように表現できると仮定する．

$$Y_f = \log(\exp(Y'_{v,f} + H(f;f_0) + g_v) + \exp(Y_{n,f} + g_n)) \quad (3)$$

ここで，$Y'_{v,f}$ 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．$Y'_{v,f}$ と $H(f;f_0)$ が足しあわされる課程は，ソースフィルタモデルの近似的表現である．さらに，$Y_{n,f}$ 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．式 (3) は，これは，声道の駆動音源に歌声のスペクトル包絡がたたみ込まれて歌声が生成され，それに伴奏が重畳したものが観測されるという伴奏を含む歌声の生成過程を，パワースペクトルの加法性を仮定した上で，確率モデルで表現したものである．. 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混合比率を変化させることができるただし，F_0 と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 $\mu'_{v,f}$, $\sigma^2_{v,f}$, $\mu_{n,f}$, $\sigma^2_{n,f}$ の推定方法は次節で述べるため，本節では既知のものとし，確率変数 Y_f はパラメータ (g_v, g_n) にのみ依存するものとする．図 2 に，以上の観測スペクトルの生成過程のまとめを示す．

W-PST 法 [6] では，$Y'_{v,f}$ と $Y_{n,f}$ が，$Y_{v,f} \sim \mathcal{N}(\mu_{v,f}, \sigma^2_{v,f})$，$Y_{n,f} \sim \mathcal{N}(\mu_{n,f}, \sigma^2_{n,f})$ のように，（対数周波数軸上で）正規分布に従うと仮定する．ここで，$\mathcal{N}(\mu, \sigma^2)$ は，平均 μ，分散 σ^2 の正規分布を表す．この場合，式 (3) 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，1 次のテーラー展開を用いて近似計算する．スペクトルが観測された時，そのスペクトルを最もよく表現する重みパラメータ (g_v, g_n) は，BFGS（Broyden-Fletcher-Goldfarb-Shanno）公式に基づく準ニュートン法を使用して推定することが可能である．これにより，確率的スペクトルテンプレート Y_f の確率密度関数を計算することが可能になる．以降の説明では，便宜的にこの確率密度関数を $p_f(y; g_v, g_n)$ と記す．

2.2.1 ウェーブレット変換に基づく声質の変換

以上により，$y(f)$ を最もよく表現する重み g_v と g_n の値が推定でき，その時の合成後のスペクトルテンプレートの確率密度関数 $p_f(y; g_v, g_n)$ が計算できる．次に，変換元の歌声包絡テンプレートのパラメータ $\mu'_{v,f}$ と $\sigma^2_{v,f}$ を変換先の歌声包絡テンプレート $\hat{\mu}'_{v,f}$ と $\hat{\sigma}^2_{v,f}$ と置き換えて，変換先のスペクトルテンプレート $\hat{p_f}(y; g_v, g_n)$ を計算する．スペクトル $y(f)$ を新しいスペクトル $\hat{y(f)}$ へ，下記の式により変換する．

$$\hat{y}(f) = y(f) + \zeta(f) \quad (4)$$

$$\zeta(f) = E_y[\hat{p}_f(y; g_v, g_n)] - E_y[p_f(y; g_v, g_n)] \quad (5)$$

ただし，$E[\cdot]$ は期待値を表す．$\zeta(f)$ はフィルターの役割を果たす関数で，元のスペクトルを変換先の歌手の歌声に変換するために操作が必要な周波数帯域とその操作量を表している．また，歌声の音量を調整したい場合は，パラメータ $\hat{g}_v$ を増減させることで実現できる．. 以上により変換後のスペクトル $\hat{y}(f)$ を得ることができる．

最後に，得られたスペクトルを逆ウェーブレット変換して，変換後の音響信号を得る．前述のように，計算時間の削減のため，式 (4) の計算は 10ms のフレーム毎に行われるので，上述の $\hat{y}(f)$ はその他の b の値では計算されない．そこで，隣り合うフレーム間の $\zeta(f)$ の値を線形補間することで，全ての b について $\zeta(f)$ を計算し，式 (4) により $\hat{y}(f)$ を計算する．時刻 b におけるフィルターを $\zeta(b,a)$ と書くと，スペクトルの位相は元のものを使うので，変換後のウェーブレットスペクトログラム $\hat{W}(b,a)$ は

$$\hat{W}(b,a) = \frac{W(b,a)}{|W(b,a)|}(|W(b,a)| + \zeta(b,a)) \quad (6)$$

で表される．ウェーブレットスペクトログラムを時間信号 $\hat{x}(t)$ に変換する逆連続ウェーブレット変換（ICWT）は，次式で定義される．

$$\hat{x}(f) = \frac{1}{C_\Psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{W}(b,a)\frac{1}{\sqrt{|a|}}\Psi\left(\frac{t-b}{a}\right)\frac{1}{a^2}dadb \quad (7)$$

ただし，C_Ψ は定数であるが，全ての時刻で同じ値をとるため厳密に計算する必要はない．

3 歌声包絡テンプレートの推定

前章では，変換前の歌声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，および変換先の歌声の歌声包絡テンプレートが与えられているという条件で，歌声変換手法について議論した．本節では，それらのテンプレートの具体的な構成方法と，テンプレートを入力音響信号から推定する手法を述べる．

スペクトルテンプレートを表現するモデルとして，文献 [6] と同様に，各回帰要素として線形回帰を使用した混合回帰モデル [7] を導入する．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置かず，学習データのみに基づいてスペクトル包絡を推定する．このモデルの未知パラメータは，EM（Expectation and Maximization）法により推定することが可能である．

3.1 単独歌唱からのテンプレート推定

単独歌唱の音響信号が与えられている場合は，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は，個別に推定する．歌声包絡テンプレートは，各母音毎に独立に推定され，例えば母音/a/のテンプレートを推定する際は，学習データ中の/a/のラベルが付与されているフレームのみを用いて推定する．ノイズスペクトルテンプレートは全体で 1 つが推定される．現在の実装では，ノイズスペクトルテンプレートの推定には，歌声を含まない伴奏のみの音響信号（カラオケトラック）を使用している．

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，

$$s_i=\{(f_{i,1},y_{i,1}),\ldots(f_{i,h},y_{i,h}),\ldots(f_{i,H_i},y_{i,H_i})\} \tag{8}$$

として表されるとする．この時，歌声包絡テンプレートを表す混合回帰モデルのパラメータ集合を θ_v とし，パラメータ θ の混合回帰モデルの周波数 f における平均と分散を，それぞれ $\mu_{v,f}(\theta_v)$ と $\sigma^2_{v,f}(\theta_v)$ とすると．最大化したい尤度関数は，次式で表される．

$$\sum_i^I\sum_h^{H_i}\log\mathcal{N}(y_{i,h}+k_i;\mu_{v,f}(\theta_v),\sigma^2_{v,f}(\theta_f)) \tag{9}$$

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．

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．

3.2 混合音からのテンプレート推定

混合音からテンプレートを推定する場合は，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を同時に推定する必要がある．I 個の観測スペクトル $y_1(f),\cdots,y_i(f),\cdots,y_I(f)$ を観測したとする．推定すべき歌声テンプレートのパラメータを θ_v とし，ノイズテンプレートのパラメータを θ_n とする．i 番目のスペクトルにおける駆動音源関数を加えた後の歌声スペクトルテンプレートは，$\mu_{v,f,i}=\mu'_{v,f}+H(f;f_0(i))$ と表すことができる．ただし，i 番目の観測スペクトルの F0 である $f_0(i)$ は全ての i について既知であるとする．

前章では，対数正規分布の加算を 1 次のテイラー展開を用いて近似計算したが，テンプレートの形状を推定する目的では式が複雑になり不向きである．そこで本節では，対数正規分布の加算を定義に従って厳密に計算した後，パラメータを近似的に推定するというアプローチをとる．合成後のスペクトルテンプレートの確率密度関数を $p_{i,f}(y;\theta_v,\theta_n,g_{i,v},g_{i,n})$ [1]と書くと，目的関数 L は，

$$L=\int\sum_{i=1}^I\log p_{i,f}(y;\theta_v,\theta_n,g_{i,v},g_{i,n})df \tag{10}$$

$$=\int\sum_{i=1}^I\log\Bigg(\int_{-\infty}^{y_i(f)}\mathcal{N}(\log(\exp(y_i(f))-\exp(U));\mu_{v,f,i}+g_{i,v},\sigma^2_{v,f})\mathcal{N}(U;\mu_{n,f}+g_{i,n},\sigma^2_{n,f})\frac{\exp(y_i(f))}{\exp(y_i(f))-\exp(U)}dU\Bigg)df \tag{11}$$

と表現される．ここで，$g_{i,v}$ と $g_{i,n}$ は，3.1 節の k_i と同様で，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．また，本節では，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 SIR を調整する役割も持っている．実際の実装では，連続ウェーブレット変換は周波数軸に対して離散的に計算しているため，f に関する積分は和の演算で置き換えられる．

ここで推定すべきパラメータは $\{g_{i,v},g_{i,n},\theta_v,\theta_n\}$ である．これらのパラメータを全て同時に最適化するのは困難であるので，逐次的に最適化する．まず，$g_{i,n}$ と θ_n を固定して，$g_{i,v}$ と θ_v の最適化と，$g_{i,v}$ と θ_v を固定して，$g_{i,n}$ と θ_n の最適化を交互に繰り返すことを考える．

まず，$g_{i,n}$ と θ_n を固定して考えると，式 (11) の和の内部は期待値の計算と考えることができる．そこで，U を期待値の計算をサンプリングにより和の計算で近似することにより，$g_{i,v}$ と θ_v の近似的な最適化を可能にする．具体的には，正規分布 $\mathcal{N}(U;\mu_{n,f}+g_{i,n},\sigma^2_{n,f})$ を $U=y_i(f)$ で切断した，上に有界な単一切断正規分布からそれぞれの i，f について R 個ずつのサンプル $(U_{i,1,f},\cdots,U_{i,r,f},\cdots,U_{i,R,f})$ をサンプリングしたとき，目的関数 L は，

$$L\approx\int\sum_{i=1}^I\log\sum_{r=1}^R\pi_{i,r,f}\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r,f}));\mu_{v,f,i}+g_{i,v},\sigma^2_{v,f}) \tag{12}$$

$$\pi_{i,r,f}=\frac{\exp(y_i(f))}{(\exp(y_i(f))-\exp(U_{i,r,f}))\,C_{y_i(f),i,f}R} \tag{13}$$

$$C_{y,i,f}=\int_{-\infty}^{y}\mathcal{N}(U;\mu_{n,f}+g_{i,n},\sigma^2_{n,f})dU \tag{14}$$

と近似できる．ここで，$g_{i,n}$ と θ_n を固定すると，$\pi_{i,r,f}$ と $\log(\exp(y_i(f))-\exp(U_{i,r,f}))$ は定数となるため，式 (12) を用いて，$g_{i,v}$ と θ_v を最適化できる．

しかし，式 (12) は和の対数の形をしているため，未だ直接の最適化が困難である．そこで，EM アルゴリズムに似た反復法によって，式 (12) を反復的に最適化する．便宜的に，推定したいパラメータを $\lambda=\{g_{i,v},\theta_v\}$ と書く．また，一回前の反復におけるパラメータの推定値を λ' と置く．まず，次式で表される変数 $z_{i,r,f}$ を導入し，

$$z_{i,r,f}=\frac{\pi_{i,r,f}\psi_{i,r,f}}{\sum_{r'=1}^R\pi_{i,r',f}\psi i,r',f} \tag{15}$$

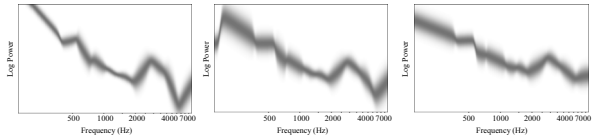
$$\psi_{i,r,f}=\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r,f}));\mu_{v,f,i}+g_{i,v},\sigma^2_{v,f}) \tag{16}$$

と置き，λ' を用いた計算した $z_{i,r,f}$ を $z'_{i,r,f}$ と書く．このとき，$z_{i,r,f}$ を固定し新たな目的関数 $Q(\lambda|\lambda')$

$$Q(\lambda|\lambda')=\int\sum_{i=1}^I\sum_{r=1}^R z'_{i,r,f}\log\pi_{i,r,f}\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r,f}));\mu_{v,f,i}+g_{i,v},\sigma^2_{v,f})df \tag{17}$$

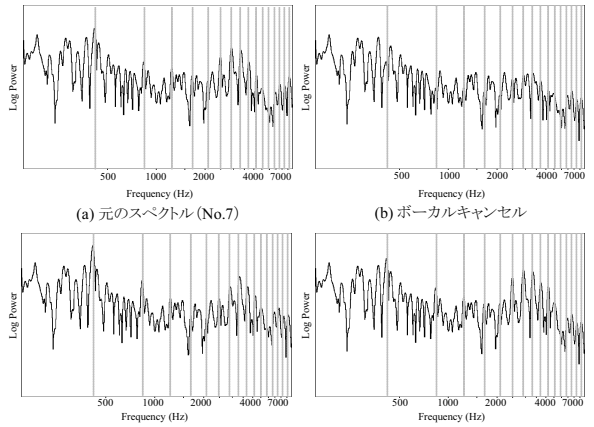
を λ に関して最適化する操作と，最適化された λ を用いて $z_{i,r,f}$ を再計算する操作を反復すると真の目的関数 L が最大化できる．式 (17) をよく見ると，$\pi_{i,r,f}$ は最適化に無関係であることがわかり，無視することができる．そのとき，Q は定数項 z の存在を除くと，式 (9) と同様の形式をしていることがわかる．すなわち，Q は 3.1 節で述べた単独歌唱からのテンプレート推定の場合と同様に最適化できることがわかる．

[1] 2.2 節と異なり，観測するスペクトルの番号 i ごとに確率密度関数の形状が異なるので，添字 i を追加している．



(a) 3.2節の手法による単独歌唱からの推定結果　(b) 提案法による推定結果　(c) 混合音から抽出した調波構造からの3.2節の手法による推定結果

Fig. 3. 混合音からのスペクトル包絡推定の例.



(a) 元のスペクトル (No.7)　(b) ボーカルキャンセル

(c) No. 13 の歌手に変換　(d) No. 20 の歌手に変換

Fig. 4. 声質変換によるスペクトル変化の例. (a) 元の No.7 のスペクトルに対して, (b) ボーカルキャンセル, (c)No.13 の歌手の声に変換, (d)No.20 の歌手の声に変換の 3 種類の処理をした場合のスペクトルを図示する. 図中の点線はスペクトルに含まれる基本周波数（約 490Hz）とその倍音周波数を表している.

4　実装

上記の技術を用い, 混合音中の歌声の声質変換を実装した. 声質の変換は正解が存在しない操作であり, 定量的な評価が困難であるので, ここでは声質を変換した場合の実験結果例をいくつか紹介する. 被験者実験等による評価実験を行うことは今後の課題となる.

まず, 混合音からのスペクトル包絡の推定の実行例を示す. 図 3 は,「RWC 研究用音楽データベース : ポピュラー音楽（RWC-MDB-P-2001）」[8] の No.7 の楽曲について, 歌声のスペクトル包絡を, 単独歌唱から 3.1 節の手法を用いた推定したもの（図 3(a)）, 混合音から 3.2 節の手法を用いて推定したもの（図 3(b)）, 混合音から抽出した調波構造を用いて 3.2 節の手法により推定したもの（図 3(c)）を図示している. (a) は単独歌唱から推定しているので理想的な推定結果と考えることができ, 提案法の推定結果 (b) がどれだけ (a) に近いかが問題となる. (c) は, 伴奏音の影響を考慮せず, 伴奏音が重畳した状態から推定した場合である. 図 3 から見てとれるように, (b) では全体に分散が大きくなる傾向や広域のパワーの弱い部分で歪みが増える傾向はあるものの, (a) に近い推定結果が得られていることがわかる. 一方, (c) では, 伴奏音の影響により, (a) と比較してスペクトルが大きく歪んでいることが見てとれる. これにより, 提案法が伴奏音の影響を低減できることがわかる.

次に, 声質変換によるスペクトル変化の実例を示す. 図 4 は, No.7 の楽曲（図 4(a)）の声質を, ボーカルをキャンセルした場合（図 4(b)）, No.13 の歌手の声に変換した場合（図 4(c)）, No.20 の歌手の声に変換した場合（図 4(d)）のスペクトルの変化である. なお, これらの使用した楽曲は全て女性で, 異なる歌手のものである. また, 1 楽曲あたり各母音が 2000 〜5000 フレーム程度含まれている. 図中のスペクトルには音素/i/の音が含まれている. なお, ボーカルキャンセルとは, 2.2.1 節の $\hat{g}_v$ を $-\infty$ に設定して声質を変換した場合であり, 声質を変換するのではなく歌声の音量を下げる変換に相当する. 図より, 伴奏音に起因する周波数成分は変化していないが, 400Hz 付近のピークや, 2500〜4500Hz 付近のピークなど, 歌声の周波数成分の形状が変化していることがわかる. 特にボーカルをキャンセルした場合は, 2500〜4500Hz 付近のピークが顕著に無くなっている.

ここで図示した以外にも, いくつかの歌手の組に対して変換を実行した. 聴感上, ボーカルキャンセルに関しては, わずかに歌声が残っているものの, 伴奏音の音質には影響を与えずに, 歌声の音量を低減できていた. 声質変換に関しては, 主観的な印象では, 変換後もわずかに元の歌手の特徴が残りながらも, 変換先の歌手の特徴を持った声に変換されているように聞こえた. 一方で, 異なる性別の歌手の声に変換する場合や, 歌声の音量を大きく増加させた場合に, 不自然な音声になることがあった. これは, 元のスペクトルで伴奏音に埋もれてしまっている周波数帯域を無理に増大させたことにより, 位相が不自然になったためだと考えられる.

5　おわりに

本稿では, 混合音中の歌声の声質変換を実現する手法について述べた. 具体的には, W-PST 法 [6] を応用して, 歌声のみの周波数成分のみを操作することを可能にした. さらに, 混合音の音響信号中の歌声のスペクトル包絡推定手法を開発することで, 歌声と伴奏と混在した状態で提供される一般の音楽音響信号に対して適用可能にした. 本技術を実装し, 実際にポピュラー音楽に対して実行することで, 提案法により正しく声質が変換されることを確認した. 今後の課題は, 被験者を用いた評価実験を行い, 提案法の性能を評価することである. また, 本稿では歌声の音素と F0 のラベルが付与されていることを仮定し, 母音に対してのみ処理をすることで, 声質が変換できることを確認した. 今後のさらなる性能向上のためには, その仮定をなくし, 全ての音素に対して処理をするために本手法を拡張していく予定である. 本研究の一部は CrestMuse プロジェクト（JST CREST）の支援を受けた.

6　参考文献

[1] Stylianou *et al.*: Continuous probabilistic transform for voice conversion, *IEEE Trans. Speech and Audio Processing*, 131–142 (1998).

[2] Mouchtaris *et al.*: Nonparallel training for voice conversion based on a parameter adaptation approach, *IEEE Trans. Audio, Speech and Language Processing*, **14**, 952–963 (2006).

[3] Toda *et al.*: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Trans. Audio, Speech and Language Processing*, **15**, 2222–2235 (2007).

[4] 河原 他: モーフィングに基づく歌唱デザインインタフェースの提案と初期検討, 情処論, **48**, 3637–3648 (2007).

[5] 大西 他: 一般逆行列を用いた母音情報に基づく声質変換法について, 信学技報, No. 282, pp. 75–80 (2007).

[6] 藤原 他: 多重奏中の歌声の基本周波数と母音音素の同時推定手法, 情処論, **51**, (2010).

[7] Jacobs *et al.*: Adaptive mixtures of local experts, *Neural Computation*, **3**, 79–87 (1991).

[8] 後藤 他: 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, 情処論, **45**, 728–738 (2004).