Vol.2010-MUS-86 No.7
2010/7/28

混合音中の歌声スペクトル包絡推定に基づく歌声の声質変換手法

藤原弘将[†1]　後藤真孝[†1]

本稿では，混合音中の歌声の声質を，別の歌手の声質に変換する手法について述べる．従来の声質変換は単独歌唱のみを対象としていた。本研究では，藤原ら[1)]によって提案された W-PST 法を応用する．W-PST 法によりスペクトル中の歌声が優勢な周波数成分を同定できる。本稿では、まず混合音のスペクトルに対して、W-PST 法で推定された歌声が優勢な周波数成分を操作することで、歌手の声質を他の歌手のものに変換する手法について述べる。次に、W-PST 法の実行に必要なスペクトル包絡を、伴奏が混在した歌声から推定する手法について述べる。本手法を実装し，実際の楽曲に対して適用したところ，歌声の声質が変換できることを確認した．

Singing voice conversion method by using spectral envelope of singing voice estimated from polyphonic music

Hiromasa Fujihara[†1] and Masataka Goto[†1]

This paper describes a singing voice conversion method that can deal with singing voices in polyphonic music. Conventional voice conversion methods only deal with monophonic singing voices. In this paper, we utilize the W-PST method proposed by Fujihara *et al.*[1)], which can identify the frequency components of a singing voice in a polyphonic spectrum. We first describe our method of converting the vocal timbres of singing voices to those of other singers by maniplulating only frequency components of singing voices identified by the W-PST method. Since the W-PST method requires spectral envelopes of the singing voices, we then describe a method of estimating them from polyphonic music. We applied our method to actual musical audio signals and confirmed that it was able to convert the vocal timbre of the singing voices in polyphonic music.

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．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。

近年，能動的音楽鑑賞インタフェース[2)]と呼ばれる，音楽を自分好みに操作しながら，より能動的に音楽を鑑賞するための技術とインタフェースが提案されている．例えば，吉井らによる Drumix[3)] は，ドラムの音量調整とドラムのパターンを置き換えが可能で，糸山らによる Instrument Equalizer[4)] では，各パートごとの音量を自由に操作しながら可能であった．本研究の技術はそのような能動的音楽鑑賞を歌声に対して実現する技術として位置づけることができ，楽曲の歌手の声質を自分好みの歌手の声に置き換えて鑑賞することが可能になる．

また，歌声合成技術の発達や Web 上の動画共有サイトの発達により，一般ユーザーが音楽を作成するようになってきており，それを支援するための製品や技術が登場している．例として，Bonada らの素片連結型歌声合成技術[5)] に基づく YAMAHA 社の歌声合成ソフトウェア VOCALOID[6)] や、酒向らの HMM 歌声合成技術[7)] に基づく歌声合成ウェブサービス Sinsy[8)] などがある。このような技術により，ユーザーは多くの人とコラボレーションしながら、より手軽に楽曲を制作し、作品を発表することができるようになっている．本研究の技術は，ユーザーが既存の伴奏を含む楽曲の声質を別の歌手や自分の声に変えることができるため，新たな音楽制作ツールとして使用できる可能性も秘めている．

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[9)–11)]，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[12)] 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

†1 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

Vol.2010-MUS-86 No.7
2010/7/28

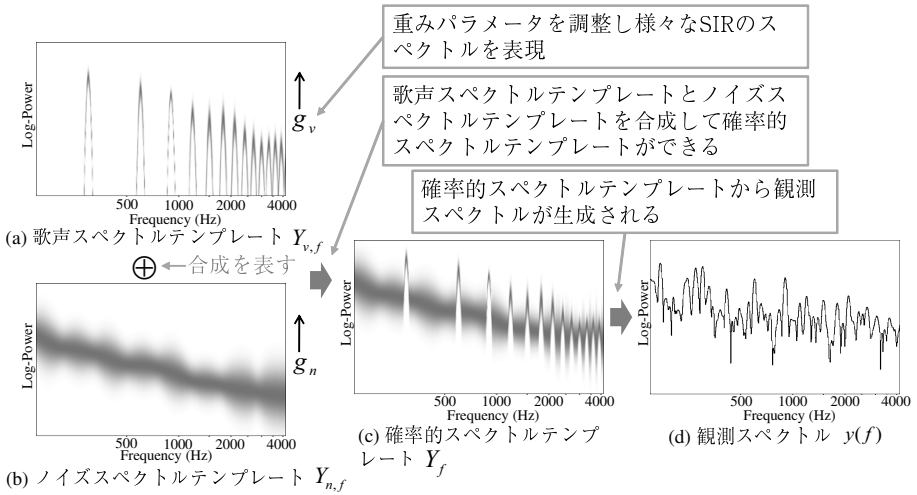


図 1 観測スペクトルの生成過程[1)]．図の濃淡は確率密度を表現する．重みパラメータ g_v と g_n を調整することで，様々な SIR のスペクトルを表現できる．

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[13)]．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．なぜなら，歌声以外の音の音質を劣化させずに，歌声のみの音質を変化させる必要があるからである．そこで本研究は，藤原ら[1)] 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．一方，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．

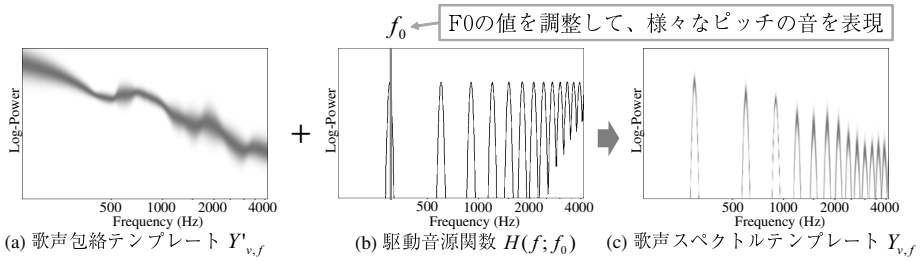


図 2 歌声スペクトルテンプレートの例[1)]．歌声包絡テンプレートと駆動音源関数から生成される．

2. W-PST 法に基づく混合音中の歌声の声質変換

本研究では，混合音中の歌声の声質変換を実現するために，藤原ら[1)] によって提案された W-PST 法を応用する．W-PST 法とは，連続ウェーブレット変換（CWT）によって得られた観測スペクトルを歌声と歌声以外の音（ノイズ）が混ざった状態としてモデル化し，音素と F0 を推定する手法である．本研究は，W-PST 法は歌声とノイズの SIR（Signal-to-Interference Ratio）を推定するため，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できることを利用し，混合音中の歌声の声質変換に応用する．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．本稿では，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 1) 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。

2.1 概　　要

W-PST 法[1)] では，**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**確率的スペクトルテンプレート**と呼ばれる確率分布の集合から生成されると仮定する．さらに，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**歌声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**ノイズスペクトルテンプレート**（図 1 (b)）の加算で表現する．つまり，観測スペクトルを生成する音源を，

Vol.2010-MUS-86 No.7
2010/7/28

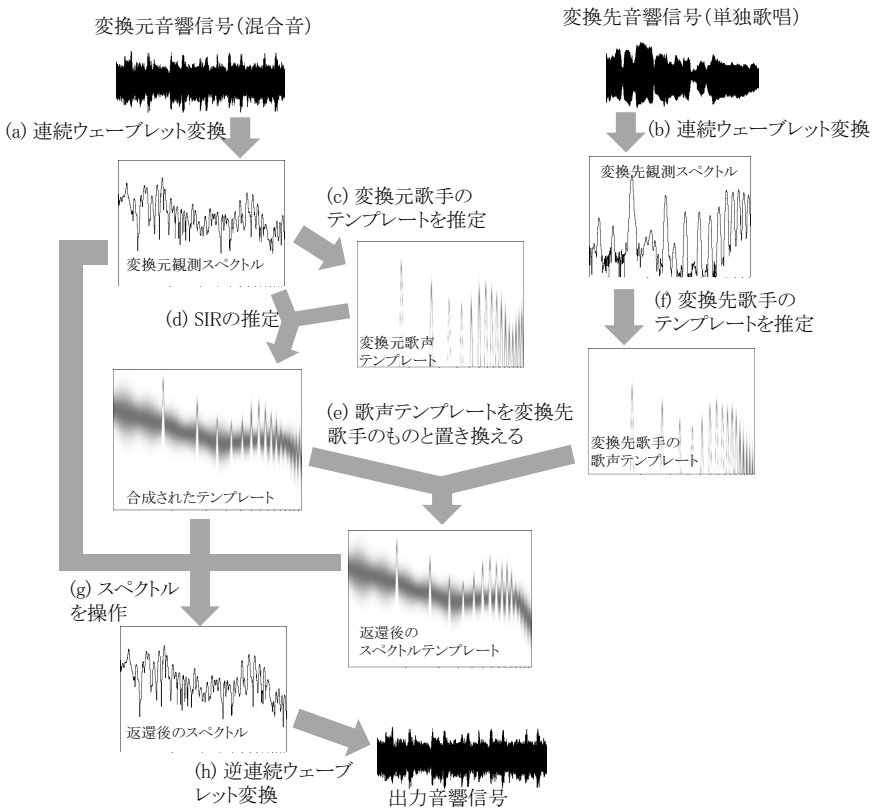


図 3 声質変換処理の概要.

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．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**（**図 2** (a)）と**駆動音源関数**（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る．

図 3 に W-PST 法を用いた声質変換の概要を示す．まず入力として変換元音響信号と変換先音響信号を取る。変換元音響信号は、市販 CD 等の実世界の音楽音響信号で、通常歌声と共に伴奏音が含まれている。一方、変換先音響信号は声質を変える目標となる歌手の音響信号で、本研究では単独歌唱であることを仮定する。前処理としてこれらの音響信号から CWT によりスペクトログラムを計算しておく（図 3 (a), (b)）。本研究では、まず W-PST 法により変換元の観測スペクトルを歌声スペクトルテンプレートとノイズスペクトルテンプレートの合成で表現することを目指す。これにより、変換元スペクトルの周波数成分ごとに、歌声が優勢なのか、伴奏成分が優勢なのかを同定することができる。そのためには、まず変換元の観測スペクトルを表すのに最適な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変換元の観測スペクトル自体から推定する必要がある（図 3 (c)）。なぜなら、変換元スペクトル中の歌手の声質は事前に単独歌唱の学習データとして準備できないからである。ここで、3.3 節で述べる混合音からのスペクトル包絡推定手法を使用する。そして、推定されたスペクトル包絡を用いて、変換元スペクトルを最もよく表現する SIR の値を、W-PST 法により計算する（図 3 (d)）。

次に、その変換元歌手をモデル化した歌声スペクトルテンプレートを、変換先の歌手をモデル化した歌声スペクトルテンプレートに置き換える（図 3 (e)）。ここで変換先の歌手の歌声スペクトルテンプレートは、単独歌唱の変換先音響信号から文献[1)] で提案された手法により推定する（図 3 (f)）。これにより、スペクトルの各周波数ビン（離散的に計算された周波数成分）ごとに、変換元の歌手から変換先の歌手へ声質変換することで、どの程度パワーを変化させる必要があるかが計算できる。ここで、ノイズスペクトルテンプレートとしては共通のものを使用しているため、歌声が優勢でない周波数帯域はパワーを変化させる必要が無くなり、結果として伴奏音の音質は保存される。そして、変換元のスペクトルの各周波数ビンのパワーを実際に変化させることで、変換先の歌声の声に変換された変換後のスペクトルを得ることができる（図 3 (g)）。最後に，変換後のスペクトルに対して逆連続ウェーブレット変換（ICWT）をかけることで音響信号を再合成する（図 3 (h)）．このとき，位相は元の観測スペクトルのものをそのまま利用する．

2.2 定　式　化

前述の手法の具体的な定式化は下記のようになる．

2.2.1 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 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．

Vol.2010-MUS-86 No.7
2010/7/28

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

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

ただし，$\overline{\Psi(\cdot)}$ は，$\Psi(\cdot)$ の共役複素数を表す現在の実装では、σ^2 を $3.75[ms]$ に、ω_0 を 1 に設定している。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．また，式 (1) の積分を無限の範囲で計算することは不可能なので，ウェーブレット $\Psi(t)$ の区間を適当に切り詰めて計算する．本稿の実験では，$-3\sigma < t < 3\sigma$ の範囲で計算をした．また，実装の際には，式 (1) は a の値を固定すると畳み込み演算であることを利用して，高速フーリエ変換 (FFT) を用いた畳み込みの高速計算法を利用した．

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 $y(f)$ を

$$y(f) = \log\left(|W(b,a)|\right) \tag{3}$$

と表記する．ただし，a と f には

$$f = \log \frac{2\pi a}{w_0} \tag{4}$$

という関係がある．

2.2.2 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 は次式により 2 つの異なるスペクトルテンプレート $Y_{v,f}$ と $Y_{n,f}$ に分割できると仮定する．

$$Y_f = \log(\exp(Y_{v,f} + g_v) + \exp(Y_{n,f} + g_n)) \tag{5}$$

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 (5) においては，パワー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．

$$Y_{v,f} \sim \mathcal{N}(\mu_{v,f}, \sigma^2_{v,f}) \tag{6}$$

$$Y_{n,f} \sim \mathcal{N}(\mu_{n,f}, \sigma^2_{n,f}) \tag{7}$$

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，スペクトル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 節で述べたように，これはソースフィルタモデルの近似的表現である．

$$Y_{v,f} = Y'_{v,f} + H(f; f_0) \tag{8}$$

$$\sim \mathcal{N}(\mu'_{v,f} + H(f; f_0), \sigma^2_{v,f}) \tag{9}$$

$$H(f; f_0) = \log\left(\sum_h \exp(-(\log f_0 + \log h - \log f)^2 / 2\theta_H^2)\right) \tag{10}$$

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．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する。また、現在の実装では、θ_H^2 は、15 cent に設定している。

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．

$$Y_f = \log(\exp(Y'_{v,f} + H(f; f_0) + g_v) + \exp(Y_{n,f} + g_n)) \tag{11}$$

確率変数 Y_f はパラメータ (g_v, g_n) に依存する．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 $p_f(y; g_v, g_n)$ と記す．

2.2.3 スペクトルテンプレートの加算の近似

式 (11) 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 $l(x_1, x_2) = \log(\exp(x_1) + \exp(x_2))$ 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

$$l(x_1, x_2) \approx \frac{\exp\left(\mu'_{v,f} + H(f; f_0) + g_v\right)}{\exp\left(\mu'_{v,f} + H(f; f_0) + g_v\right) + \exp\left(\mu_{n,f} + g_n\right)} x_1 + \frac{\exp\left(\mu_{n,f} + g_n\right)}{\exp\left(\mu'_{v,f} + H(f; f_0) + g_v\right) + \exp\left(\mu_{n,f} + g_n\right)} x_2 + C \tag{12}$$

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 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，$Y_f = l(Y'_{v,f} + H(f; f_0) + g_v, Y_{n,f} + g_n)$ が従う確率密度関数 $p_f(y; g_v, g_n)$ は，

$$p_f(y;g_v,g_n) \approx \mathcal{N}(y;\mu_f(\theta_v,\theta_n),\sigma_f^2(\theta_v,\theta_n)) \tag{13}$$

$$\mu_f(g_v,g_n) = \log(\exp(\mu'_{v,f}+H(f;f_0)+g_v)+\exp(\mu_{n,f}+g_n)) \tag{14}$$

$$\sigma_f^2(g_v,g_n) = \frac{(\exp(\mu'_{v,f}+H(f;f_0)+g_v))^2\sigma_{v,f}^2+(\exp(\mu_{n,f}+g_n))^2\sigma_{n,f}^2}{(\exp(\mu'_{v,f}+H(f;f_0)+g_v)+\exp(\mu_{n,f}+g_n))^2} \tag{15}$$

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．

2.2.4 準ニュートン法によるパラメータ最適化

SIR を表すパラメータ (g_v,g_n) 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 $Q(g_v,g_n)$ は，

$$Q(g_v,g_n) = -\sum_f \log\mathcal{N}(y(f);u_f(g_v,g_n),\sigma_f^2(g_v,g_n)) \tag{16}$$

で表される．ただし，$y(f)$ は観測スペクトルである．

2.2.5 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．

$$\hat{y}(f) = y(f)+\zeta(f) \tag{17}$$

$$\zeta(f) = E_y[\hat{p}_f(y;g_v,g_n)] - E_y[p_f(y;g_v,g_n)] \tag{18}$$

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 (17) 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 (17) 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

$$\hat{W}(b,a) = \frac{W(b,a)}{|W(b,a)|}(|W(b,a)|+\zeta(b,a)) \tag{19}$$

で表される．ただし，a と f には，式 (4) のような関係がある．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 ICWT は，次式で定義される．

$$\hat{x}(f) = \frac{1}{C_\Psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{W}(b,a)\frac{1}{\sqrt{|a|}}\Psi\left(\frac{t-b}{a}\right)\frac{1}{a^2}dadb \tag{20}$$

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．ICWT の計算においても CWT と同様で，FFT による畳み込み演算の高速計算法を用いて実装した．

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．

3.1 混合回帰モデルによるテンプレートの表現

スペクトルテンプレートを表現するモデルとして，文献 1) 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[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．混合数 M の混合回帰モデルは，$m(1,\cdots,M)$ を各線形回帰モデルのインデックスとすると、パラメータとして、それぞれの線形回帰モデルの傾き $a_{v,m}$ と切片 $b_{v,m}$、各線形回帰モデルの守備範囲を決めるゲート関数のパラメータ $\psi_{v,m}$、$\mu_{v,m}$、$\sigma^2_{v,m}$ をとる。ゲート関数としては、次式で定義される正規化ガウス関数[15] を用いた．

$$G_m(f;\psi_{v,m},\mu_{v,m},\sigma^2_{v,m}) = \frac{\psi_{v,m}\mathcal{N}(f;\mu_{v,m},\sigma^2_{v,m})}{\sum_{m'=1}^{M}\psi_{m'}\mathcal{N}(f;\mu_{v,m'},\sigma^2_{v,m'})} \tag{21}$$

ここで、$\psi_{v,m}$ は各ガウス関数の重みを決めるパラメータで、$\psi_{v,m}\geq 0$ かつ $\sum_{m=1}^{M}\psi_{v,m}=1$ である．また、$\mu_{v,m}$ と $\sigma^2_{v,m}$ は、ガウス関数の平均と分散である。このモデルでは、歌声

包絡テンプレートのパラメータである平均 $\mu'_{v,f}$ と分散 $\sigma^2_{v,f}$ は

$$\mu'_{v,f} = \sum_{m=1}^{M} G_m(f;\psi_{v,m},\mu_{v,m},\sigma^2_{v,m})(a_{v,m}f + b_{v,m}) \tag{22}$$

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

として表現する．ただし，M は混合数を表す．現在の実装では、M を 10 に設定している。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．ノイズスペクトルテンプレートについても同様で，未知パラメータを $\{\psi_{n,m},\mu_{n,m},\sigma^2_{n,m},a_{n,m},b_{n,m},\beta^2_{n,m}\}$ と置き，同様の形式で表現する．

3.2 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。

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．

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，

$$s_i = \{(f_{i,1},y_{i,1}),\ldots(f_{i,h},y_{i,h}),\ldots(f_{i,H_i},y_{i,H_i})\} \tag{24}$$

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．

$$\sum_{i}^{I}\sum_{h}^{H_i} \log\mathcal{N}(y_{i,h}+k_i;\mu_{v,f_{i,h}},\sigma^2_{v,f_{i,h}}) \tag{25}$$

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．

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．

$$k_i = \frac{\sum_{h=1}^{H_i}\dfrac{\mu_{v,f_{i,h}} - y_{i,h}}{\sigma^2_{v,f_{i,h}}}}{\sum_{h=1}^{H_i}\dfrac{1}{\sigma^2_{v,f_{i,h}}}} \tag{26}$$

Step 3 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 (現在の実装では 60Hz〜7500Hz) 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^2_m の初期値に設定し，ψ_m の初期値は $\frac{1}{M}$ とした．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．現在の実装では、Step 1 の EM 法の反復回数は 1 にし、Step 0〜3 全体の反復回数を 30 回に設定している。

3.3 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は $\theta_v = \{\psi_{v,m},\mu_{v,m},\sigma^2_{v,m},a_{v,m},b_{v,m},\beta^2_{v,m}\}$ とし，ノイズテンプレートのパラメータは $\theta_n = \{\psi_{n,m},\mu_{n,m},\sigma^2_{n,m},a_{n,m},b_{n,m},\beta^2_{n,m}\}$ 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，

$$\mu_{v,f,i} = \mu'_{v,f} + H(f;f_0(i)) \tag{27}$$

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．しかし，得られた式 (13)〜(15) は複雑な形状となり，θ_v，θ_n を最適化するのは困難である．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，

⋆1 2.2.3 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．

$$L = \int \sum_{i=1}^{I} \log p_{i,f}(y;\theta_v,\theta_n,g_{i,v},g_{i,n})df \tag{28}$$

$$= \int \sum_{i=1}^{I} \log(\int_{-\infty}^{y_i(f)} \mathcal{N}(\log(\exp(y_i(f))-\exp(U));\mu_{v,f,i}+g_{i,v},\sigma_{v,f}^2)$$
$$\mathcal{N}(U;\mu_{n,f}+g_{i,n},\sigma_{n,f}^2)\frac{\exp(y_i(f))}{\exp(y_i(f))-\exp(U)}dU)df \tag{29}$$

$$= \int \sum_{i=1}^{I} \log(\int_{-\infty}^{y_i(f)} \mathcal{N}(\log(\exp(y_i(f))-\exp(U));\mu_{n,f}+g_{i,n},\sigma_{n,f}^2)$$
$$\mathcal{N}(U;\mu_{v,f,i}+g_{i,v},\sigma_{v,f}^2)\frac{\exp(y_i(f))}{\exp(y_i(f))-\exp(U)}dU)df \tag{30}$$

と表現される．ここで，$g_{i,v}$ と $g_{i,n}$ は，3.2 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，式 (29) による $g_{i,v}$ と θ_v の最適化と，$g_{i,v}$ と θ_v を固定して，式 (30) による $g_{i,n}$ 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 (29) 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 $\mathcal{N}(U;\mu_{n,f}+g_{i,n},\sigma_{n,f}^2)$ 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，

$$L \approx \int \sum_{i=1}^{I} \log \sum_{r=1}^{R} \pi_{i,r,f}\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r,f}));\mu_{v,f,i}+g_{i,v},\sigma_{v,f}^2) \tag{31}$$

$$\pi_{i,r,f} = \frac{\exp(y_i(f))}{(\exp(y_i(f))-\exp(U_{i,r,f}))\,R\int_{-\infty}^{y_i(f)}\mathcal{N}(U;\mu_{n,f}+g_{i,n},\sigma_{n,f}^2)dU} \tag{32}$$

と近似できる．現在の実装では R の値を 300 に設定している。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 (31) を用いて，$g_{i,v}$ と θ_v を最適化できる．また，$g_{i,v}$ と θ_v を固定した場合も同様で，式 (29) からサンプリングにより式 (31) と同様の式を導出し，$g_{i,n}$ と θ_n を最適化する．

しかし，式 (31) 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 (31) 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，変数 $z_{i,r,f}$ を導入し，

$$z_{i,r,f} = \frac{\pi_{i,r,f}\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r,f}));\mu_{v,f,i}+g_{i,v},\sigma_{v,f}^2)}{\sum_{r'=1}^{R}\pi_{i,r',f}\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r',f}));\mu_{v,f,i}+g_{i,v},\sigma_{v,f}^2)} \tag{33}$$

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 $Q_1(\lambda|\lambda')$

$$Q_1(\lambda|\lambda') = \int \sum_{i=1}^{I}\sum_{r=1}^{R} z'_{i,r,f}\log\pi_{i,r,f}$$
$$\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r,f}));\mu_{v,f,i}+g_{i,v},\sigma_{v,f}^2)df \tag{34}$$

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．証明は付録を参照されたい．

式 (34) 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，関数 $Q_2(\lambda|\lambda')$

$$Q_2(\lambda|\lambda') = \int \sum_{i=1}^{I}\sum_{r=1}^{R} z'_{i,r,f}\log\mathcal{N}(\log(\exp(y_i(f))-\exp(U_{i,r,f}));\mu_{v,f,i}+g_{i,v},\sigma_{v,f}^2)df \tag{35}$$

の最適化は，$Q_1(\lambda|\lambda')$ の最適化と等価であることがわかる．さらに，Q_2 は定数項 z の存在を除くと，式 (25) と同様の形式をしていることがわかる．すなわち，Q_2 は 3.2 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．

以上をまとめるとパラメータは下記の手続きで推定される．

Step 0 $g_{i,v}=0$，$g_{i,n}=0$ 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 $g_{i,n}$ と θ_n を固定して，式 (29) の U をサンプリングする．

Step 2 サンプリングした U と現在のパラメータ $g_{i,v}$，θ_v を用いて，式 (33) の $z_{i,r,f}$ を計算する．

Step 3 計算された $z_{i,r,f}$ を用いて，式 (35) の Q_2 関数を最適化する．この最適化には 3.2 節の反復的な最適化法を利用する．

Step 4 Step 2〜3 の反復が規定回数を超えた場合は Step 5 へ，そうでない場合は Step 2 に戻る．

Step 5 $g_{i,v}$ と θ_v を固定して，式 (30) の U をサンプリングする．

Step 6 サンプリングした U と現在のパラメータ $g_{i,n}$，θ_n を用いて，式 (33) の $z_{i,r,f}$ を計算する．

Step 7 計算された $z_{i,r,f}$ を用いて，式 (35) の Q_2 関数を最適化する．この最適化には 3.2 節の反復的な最適化法を利用する．

Step 8 Step 2〜3 の反復が規定回数を超えた場合は Step 9 へ，そうでない場合は Step 6 に戻る．

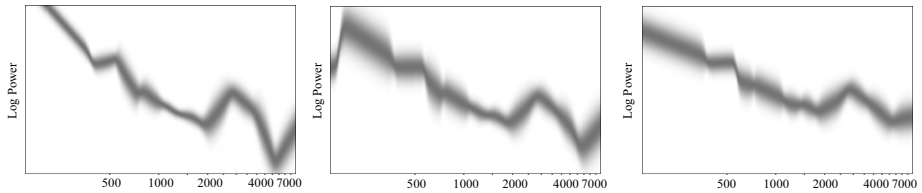
Step 9 Step 1〜8 の反復が規定回数を超えた場合は終了する．そうでない場合は Step 1 に戻る．

歌声包絡テンプレートの初期値は、今回の推定対象とは異なる歌手の単独歌唱の音響信号から、ノイズスペクトルテンプレートの初期値は，歌声の入っていない音楽音響信号（カラオケトラック）から、それぞれ 3.2 節の手法により推定したパラメータの値を使用する．

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 4 は，「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 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 3.2 節の手法を用いた推定したもの（図 4(a)），混合音から 3.3 節の手法を用いて推定したもの（図 4(b)），混合音から抽出した調波構造を用いて 3.3 節の手法により推定したもの（図 4(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 4 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。

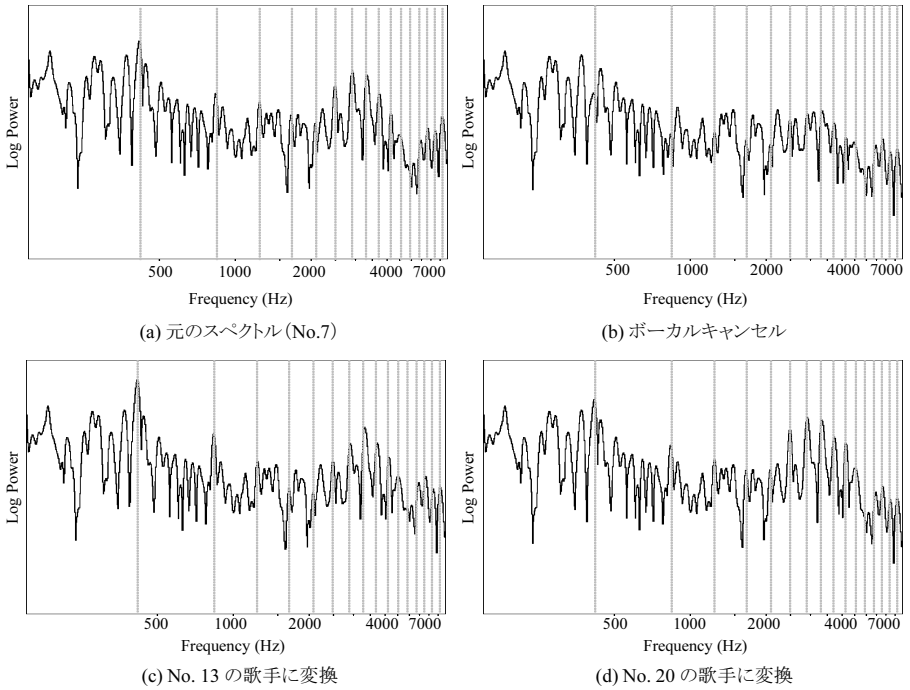


(a) 3.2節の手法による単独歌唱からの推定結果　(b) 提案法による推定結果　(c) 混合音から抽出した調波構造からの3.2節の手法による推定結果

図 4 混合音からのスペクトル包絡推定の例.「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 No.7 の楽曲から推定した/i/の音素のスペクトル包絡である．

次に，声質変換によるスペクトル変化の実例を示す．図 5 は，No.7 の楽曲（図 5(a)）の声質を，ボーカルをキャンセルした場合（図 5(b)），No.13 の歌手の声に変換した場合（図 5(c)），No.20 の歌手の声に変換した場合（図 5(d)）のスペクトルの変化である．また，図 6 は，図 5 において元の楽曲を No.20 の声質に変換した場合の，それぞれに対応するスペクトルテンプレートの例である．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，2.2.5 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，楽曲によっては，変換元の歌手の声と変換先の歌手の声が混ざったような声になる場合もあった．また，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



(a) 元のスペクトル（No.7）　(b) ボーカルキャンセル

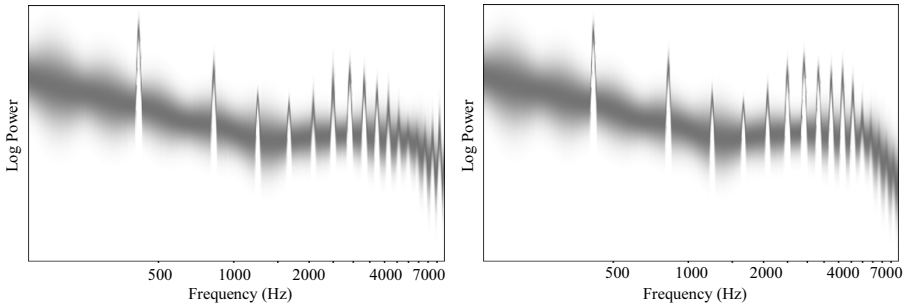
(c) No. 13 の歌手に変換　(d) No. 20 の歌手に変換

図 5 声質変換によるスペクトル変化の例.「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 No.7 の楽曲に対して，(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。

になったためだと考えられる．これに対しては，そのような周波数帯域では，歌声の周波数成分を正弦波重畳モデル等で別に再合成して足しあわせるなどの処理が有効であると考えられる．

5. ま と め

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[1] 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



(a) 元のスペクトル（No.7）を表現する合成されたスペクトルテンプレート

(b) No. 20の歌手の歌声包絡テンプレートを用いて操作されたスペクトルテンプレート

図 6 声質変換の際のテンプレートの置換の例.「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 No.7 の楽曲のスペクトルを表現する合成されたテンプレート (a) と，その歌声を No.20 のテンプレートに置換したテンプレート (b) を図示する．(a) は図 5 (a) に，(b) は図 5 (d) に対応する．

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。

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．また、本稿での実験に「RWC 研究用音楽データベース：ポピュラー音楽（RWC-MDB-P-2001）」[16] を使用した。

参 考 文 献

1) 藤原弘将，後藤真孝，奥乃　博：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，情報処理学会研究報告，Vol.2009-MUS-81 (2009).
2) Goto, M.: Active Music Listening Interfaces Based on Signal Processing, *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, pp.IV–1441–1444 (2007).
3) Yoshii, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Drumix: An Audio Player with Real-time Drum-part Rearrangement Functions for Active Music

Listening, *IPSJ Journal*, Vol.48, No.3, pp.1229–1239 (2007).
4) Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Instrument Equalizer for Query-by-Example Retrieval: Improving Sound Source Separation based on Integrated Harmonic and Inharmonic Models, *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pp.133–138 (2008).
5) Bonada, J., Celma, O., Loscos, A., Ortola, J., Serra, X., Yoshioka, Y., Kayama, H., Hisaminato, Y. and Kenmochi, H.: Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models, *Proceedings of International Computer Music Conference* (2001).
6) 剣持秀紀，大下隼人：歌声合成システム VOCALOID，情報処理学会研究報告，Vol.2007-MUS-72, pp.25–28 (2007).
7) 酒向慎司，宮島千代美，徳田恵一，北村　正：隠れマルコフモデルに基づいた歌声合成システム，情報処理学会論文誌，Vol.45, No.3, pp.719–727 (2004).
8) Sinsy - HMM-based Singing Voice Synthesis System: http://www.sinsy.jp/.
9) Stylianou, Y., Cappé, O. and Moulines, E.: Continuous probabilistic transform for voice conversion, *IEEE Transactions on Speech and Audio Processing*, No. 2, pp. 131–142 (1998).
10) Mouchtaris, A., der Spiegel, J.V. and Mueller, P.: Nonparallel training for voice conversion based on a parameter adaptation approach, *IEEE Transactions on Audio, Speech and Language Processing*, Vol.14, No.3, pp.952–963 (2006).
11) Toda, T., Black, A.W. and Tokuda, K.: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Transactions on Audio, Speech and Language Processing*, Vol.15, No.8, pp.2222–2235 (2007).
12) 河原英紀，生駒太一，森勢将雅，高橋　徹，豊田健一，片寄晴弘：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，情報処理学会論文誌，Vol.48, No.12, pp.3637–3648 (2007).
13) 大西壮登，高橋　徹，入野俊夫，河原英紀：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，電子情報通信学会技術報告，No.282, pp.75–80 (2007).
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) 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).

付録　3.3 節の反復アルゴリズムの妥当性の証明

証明．3.3 節において，式 (31) の L は，Jensen の不等式より

$$L(\lambda) = \int \sum_{i=1}^{I} \log \left(\sum_{r=1}^{R} z'_{i,r,f} \frac{\pi_{i,r,f} \mathcal{N}(x_{i,r,f}; \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f})}{z'_{i,r,f}} \right) df \tag{36}$$

$$\geq \int \sum_{i=1}^{I} \sum_{r=1}^{R} z'_{i,r,f} \log \frac{\pi_{i,r,f} \mathcal{N}(x_{i,r,f}; \mu_{v,f,i} + g_{i,v}, \sigma^2_{v,f})}{z'_{i,r,f}} df = F(\lambda|\lambda') \tag{37}$$

と変形できる．ただし，$x_{i,r,f}$ は，

$$x_{i,r,f} = \log(\exp(y_i(f)) - \exp(U_{i,r,f})) \tag{38}$$

である．このとき，

$$L(\lambda) - L(\lambda') = F(\lambda|\lambda') - F(\lambda'|\lambda') + \int \sum_{i=1}^{I} z'_{i,r,f} \log \left(\frac{z_{i,r,f}}{z'_{i,r,f}} \right) df \tag{39}$$

が成立する．右辺第三項は非負なので $F(\lambda|\lambda^{(-1)})$ の λ に関する最大化は目的関数 $L(\lambda)$ を増加させることがわかる．さらに，

$$Q_1(\lambda|\lambda') = F(\lambda|\lambda') + \int \sum_{i=1}^{I} z'_{i,r,f} \log \left(z'_{i,r,f} \right) df \tag{40}$$

と変形でき，右辺第二項は λ に無関係な項であるので，$F(\lambda|\lambda')$ の λ に関する最大化は，$Q_1(\lambda|\lambda')$ の λ に関する最大化と等価である．以上より，$Q(\lambda|\lambda')$ を最大化させることで，目的関数 $L(\lambda)$ が増加することが示される．□