社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (23)
2008－SLP－70 (23)
2008／2／9

声質制御への応用を目的とした声道断面積関数の分析

内村 佳典† 坂野 秀樹† 板倉 文忠†

† 名城大学大学院理工学研究科 〒468-5802 愛知県名古屋市天白区塩釜口 1-501
E-mail: †m0732008@ccmailg.meijo-u.ac.jp

あらまし 声質制御の新たな手法として声道断面積関数による声質制御方式を提案する．この方法は，声道の形状を表すパラメタとして知られる Kelly の音声生成モデルに基づく声道断面積関数を変化させることにより声質を制御する．本稿では，まず，この声道断面積関数を求める際に行う音源と放射の特性の除去方法に関して検討を行った．その結果，中島らによって提案された適応逆フィルタを用いる方法が適していることが分かった．また，声道断面積関数の形状はサンプリング周波数が異なる場合でも破綻せずに推定できることが明らかとなった．更に，提案法を用い，ある話者の声道断面積関数の平均値を別の話者の声道断面積関数の平均値に置き換えるという処理により，簡易な声質変換を実行するシステムを構築した．その結果，合成音声の品質及び共振周波数がターゲットのものに極めて近くなるという結果が得られた．

Analysis of the vocal tract area function aimed at manipulation of voice quality

Yoshinori UCHIMURA†, Hideki BANNO†, and Fumitada ITAKURA†

† Graduate School of Science and Technology, Meijo University Siogamaguchi 1-501, Tempaku-ku, Nagoya-shi, Aichi 468-5802 Japan
E-mail: †m0732008@ccmailg.meijo-u.ac.jp

Abstract This paper describes a new manipulation method of voice quality based on the vocal tract area function. This method can manipulate the voice quality by using modification of the vocal tract area function which represents a shape of a vocal tract simulated by the Kelly's speech production model. In the extraction of the vocal tract area function, the characteristics of sound radiation and glottal source of speech should be canceled out. Several methods have been proposed to cancel out the characteristics. We have evaluated the cancellation methods in the voice quality manipulation, and found that the method based on adaptive inverse filtering proposed by Nakajima et al. is suitable for the voice quality manipulation. We also found that the vocal tract area function is successfully extracted even if an input speech signal is digitized with a high sampling frequency. A simple system that converts the voice quality by shifting the mean of the vocal tract area function of input speech to that of target speech was constructed and confirmed to produce similar sounds to the target voice.

1. はじめに

人間の音声は，声帯の振動周期と声道の共鳴特性によって特徴付けられる．声道の共鳴特性は，声道の形状により変化する．人間は同じ発話内容でも感情や声の高さによって声道の形状が変化し，その共鳴特性も変化する．また，話者によっても声道の形状は異なっている．従って本稿では，この声道の形状に着目し声道の形状を表すパラメタである声道断面積関数 [1] [2] により声質を制御する方式を提案する．

近年，高品質な音声分析合成が可能な STRAIGHT [3] などのシステムが音声モーフィング [4] など様々な声質制御に用いられている．高品質な音声分析合成が可能となったことで声質制御方式への要求も高くなってきており，自然な声質制御が可能なことや，入力音声のサンプリング周波数が高い場合にも破綻することなく高品質な合成が可能であることなどが求められる．本稿では，様々な音声について分析を行い，提案法がこのような条件を満たす可能性があることを示す．

また，声道断面積関数を求める際の音声波形に行う音

源と放射の特性の除去方法や，提案法の話者変換への応用に関する検討を行った．

2. 声道断面積関数による声質制御方式

声質は，声道の形状が変化することによる声道の共鳴特性の変化に大きく依存している．このため提案法では，ある音声から求めた声道断面積関数を変化させることにより声質を制御する．ここでいう声道断面積関数とは，Kelly の音声生成モデルに基づくもので，声道を 図 1 のような 断面積一定の微小音響管の従属接続で表現した場合の各区間の断面積の組である．図 1 の l は微小音響

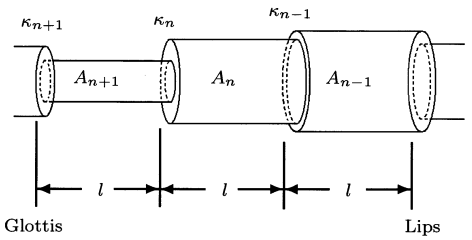

図 1　声道モデル

管の長さ，A_n は区間 n の断面積，κ_n は区間 n と区間 $n+1$ 間の反射係数を表している．図 2 に音声波形から声道断面積関数を求める手順を示す．

音声波形は音源と放射の特性を含んでいるため，声道

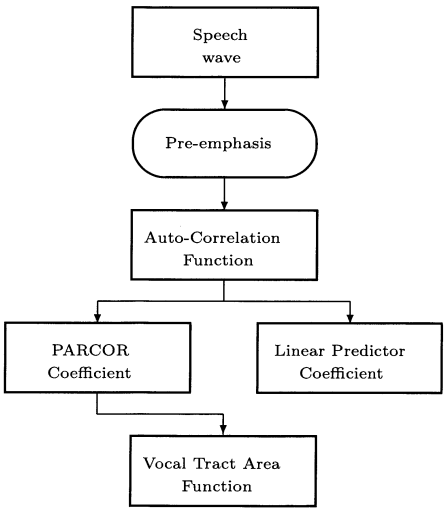

図 2　音声波形から声道断面積関数を求める手順

断面積関数を求める前にこれらの特性を除去しておく必要がある．これにはいくつかの方法がる上，この方法によって声質制御後の合成音声も変化するため，次節で詳しく述べる．音源と放射の特性の除去を行った後，その音声波形の自己相関関数を用いて PARCOR 係数を求める．ここで，反射係数 κ_n は式 (1) で表されるため，

$$\kappa_n = \frac{A_n - A_{n+1}}{A_n + A_{n+1}} \tag{1}$$

反射係数 κ_n と断面積 A_n には式 (2) の関係が成り立ち，

$$A_n = \frac{1+\kappa_n}{1-\kappa_n} A_{n+1} \tag{2}$$

また，PARCOR 係数 k_n は反射係数 κ_n に対応しているため，この PARCOR 係数 k_n から式 (3) により声道断面積関数を求める．[2]

$$\begin{aligned} A_n &= \frac{1+k_n}{1-k_n} A_{n+1} \quad n = p, p-1, \cdots, 1 \\ A_{p+1} &= 1 \end{aligned} \tag{3}$$

この声道断面積関数の値を変化させたものを声質制御後の声道断面積関数とする．

次に，この声質制御後の声道断面積関数から合成音声を作成する手順について説明する．今回は，残差信号を駆動音源とする LPC ボコーダにより合成音声を作成している．このため，声質制御後の声道断面積関数を線形予測係数に変換する必要がある．まず，声道断面積関数から PARCOR 係数に変換する．声道断面積関数は式 (1) より，式 (4) で PARCOR 係数に変換することができる．

$$k_n = \frac{A_n - A_{n+1}}{A_n + A_{n+1}} \tag{4}$$

PARCOR 係数 k_n から線形予測係数 $\alpha_n (= \alpha_n^{(p)})$ は図 3 の手順により求めることができる．この線形予測係数を用いて合成音声を作成する．図 4 に声質制御の例として，提案法により声質制御された合成音声と原音声の式 (5) [5] により得られた共振周波数の時間変化を示す．

$$F_i = \frac{\arg z_i}{2\pi} F_s \quad [\mathrm{Hz}] \tag{5}$$

ここで，F_s はサンプリング周波数，z_i は $1+\sum \alpha_i z^{-i} = 0$ の根である．この声質制御された合成音声は，次節で述べる適応逆フィルタにより音源と放射の特性の除去を行った音声波形から求めた声道断面積関数の口唇に最も近い区間の断面積の値を大きくしたものである．図 4 から，共振周波数が変化していることが確認できる．

3. 音源・放射特性の除去方法

前節でも述べたが，音声波形は音源と放射の特性を含んでいるため，声道断面積関数を求める前にこれらの特性を除去しておく必要がある．この方法として，音声波形を式 (6) のディジタルフィルタに通す方法や，

$$\begin{aligned} &H(z) = 1 - az^{-1} \\ &(\text{ただし } a \text{ の値は } 0.98 \text{ とした}) \end{aligned} \tag{6}$$

中島らによって提案された適応逆フィルタによる方法 [1]

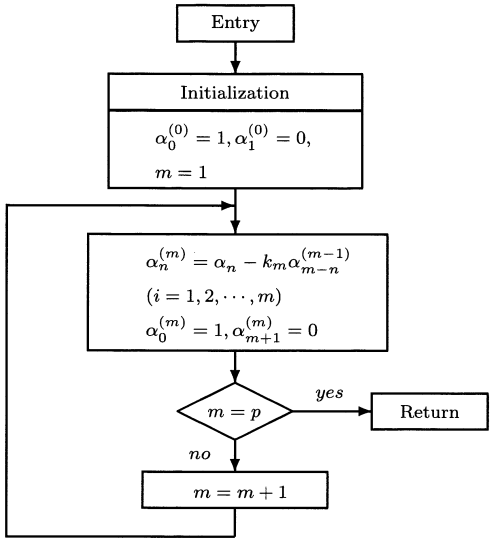

図 3　PARCOR 係数 $k_n(n=1,2,\cdots,p)$ から線形予測係数 $\alpha_n(=\alpha_n^{(p)})$ を求める手順

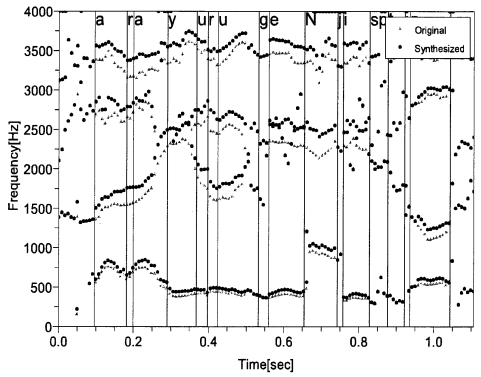

図 4　提案法による共振周波数の変化

などが知られている．前者は，音源と放射の周波数特性が，それぞれ-12dB/oct と 6dB/oct で近似できるため，差し引き 6dB/oct の高域強調を行う．後者は，声道の周波数特性はほとんど平坦で傾きをもたない [6] という前提で，分析フレームごとにその傾き特性を取り除くものである．図 5 に適応逆フィルタの構成を示す．ここで ϵ_i $(i=1,2,4)$ は音声波形の共分散行列の要素を

$$C_{jk}=\sum_{i=0}^{N-1}x_{t-j-i}x_{t-k-i} \tag{7}$$

とすると，次の 3 次代数方程式のうち $|\epsilon_i|<2$ を満たす実根で与えられる．

$$C_{22}\epsilon_i^3-C_{21}\epsilon_i^2+(4C_{02}+8C_{11})\epsilon_i-8C_{01}=0 \tag{8}$$

また，ϵ_3,ϵ_5 はそれぞれ前段までの残差波形について 2 次

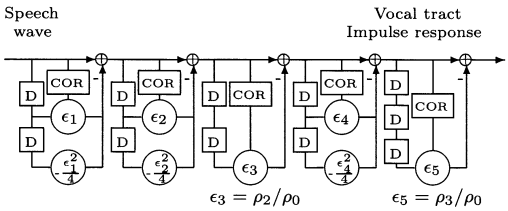

図 5　適応逆フィルタの構成．ρ_i は i 次の相関係数

及び 3 次の相関係数に対応する．

図 6 に式 (6) のディジタルフィルタを用いた場合の各音素の分析結果を示す．分析に用いた音声は，男性話者が「あいうえお」と発声したものでサンプリング周波数は 11025Hz である．図の左はフィルタにより除去される特性と，除去後の各音素の音声波形の線形予測分析によるスペクトル包絡の長時間平均スペクトルである．右は，各音素の声道断面積関数の平均値である．この横軸は口唇からの距離を表しておりこの 1 区間の長さ $l$cm は，音速を c m/s サンプリング周波数を F_s Hz とすると

$$l=100c/2F_s \quad [\text{cm}] \tag{9}$$

で与えられる．[2] 図 7 は適応逆フィルタを用いた場合の，図 6 と同様の分析結果である．

この結果から，これらの 2 つの方法は，除去される特性が異なるため得られる声道断面積関数も大きく異なることが分かる．特に，式 (6) のディジタルフィルタを用いた場合には，図 6 の/e/，/a/，/o/，/u/の声道断面積関数の口唇付近のように局部的に極めて大きな値になることがあった．このため，声道断面積関数を同じように変化させて得られる合成音声の品質は異なる．そこで，提案法に用いる高域強調の方法としてどちらが適切か予備的な比較検討を行った．ここでは，それぞれの方法により音源と放射の特性を除去した音声波形について声質制御を行った．その結果，式 (6) のフィルタによる方法の合成音声は人間の音声として不自然なものとなることがあった．このため，提案法では適応逆フィルタにより音源と放射の特性を除去することとした．なお，この適応逆フィルタは時間領域での処理であるが，STRAIGHT への適用を考え，周波数領域で同様の処理を行う方法について現在検討中である．

4. サンプリング周波数による影響

はじめにも述べたが，近年 STRAIGHT などのシステムに様々な声質制御が用いられるようになったことで，声質制御への要求水準も高くなってきている．高品質な声質制御のためには，声道断面積関数の分析結果がサンプリング周波数に依存しないことが要求される．そこで，声道断面積関数の分析結果のサンプリング周波数による影響について検討を行った．

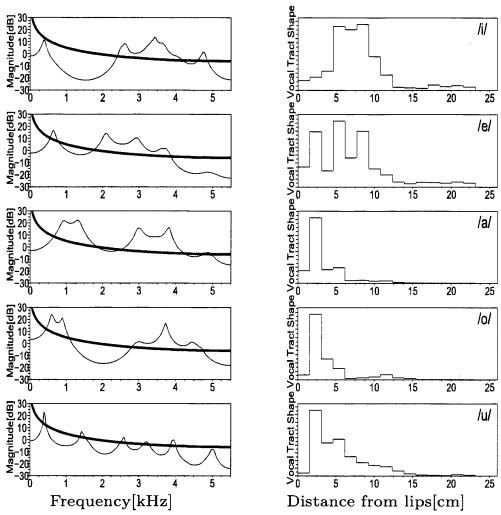

図 6　式 (6) のフィルタを用いた場合の各音素の分析結果．左：式 (6) フィルタにより除去される特性と高域強調された音声波形の線形予測分析によるスペクトル包絡の長時間スペクトル，右：道断面積関数，話者：男性，サンプリング周波数：11025Hz

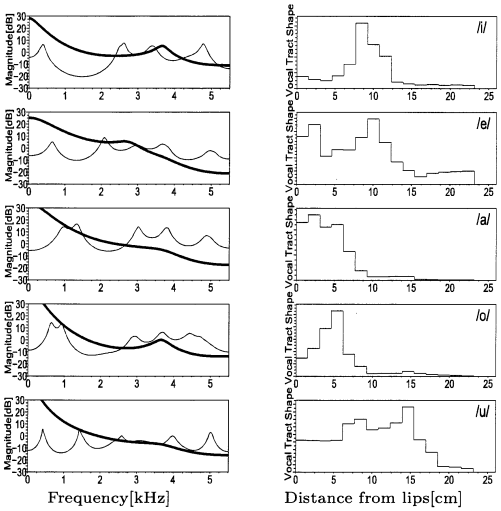

図 7　適応逆フィルタを用いた場合の各音素の分析結果．左：適応逆フィルタにより除去される特性と高域強調された音声波形の線形予測分析によるスペクトル包絡の長時間スペクトル，右：道断面積関数，話者：男性，サンプリング周波数：11025Hz

図 8 に 44100Hz で収録した音声とその音声をダウンサンプリングしたものを用いて，声道断面積関数の時間変化のサンプリング周波数による違いを比較した結果を示す．上からサンプリング周波数が 44100Hz の場合，22050Hz の場合，11050Hz の場合の分析結果である．縦軸は口唇からの距離を表しており，横軸は時間を表している．また，濃淡は声道断面積の大小を表しており，濃淡が濃い部分が声道断面積の値が大きく薄い部分が小さいことを表している．

図 8 から，サンプリング周波数が異なる場合でも，濃淡の位置はほぼ対応しているため，サンプリング周波数の高低に関係なく同じ枠組みで声質制御可能であると考えられる．ここで，/u/や無音区間で濃淡のパターンが異なっている箇所も見られるが，これは，図 5 に示した適応逆フィルタの構成が，主に男性話者に対して，標本化周波数 12kHz 程度の場合を対象としていることが原因と考えられる．これが声質制御後の合成音声にどの程度影響を及ぼすかについては，今後，検討を行う予定である．

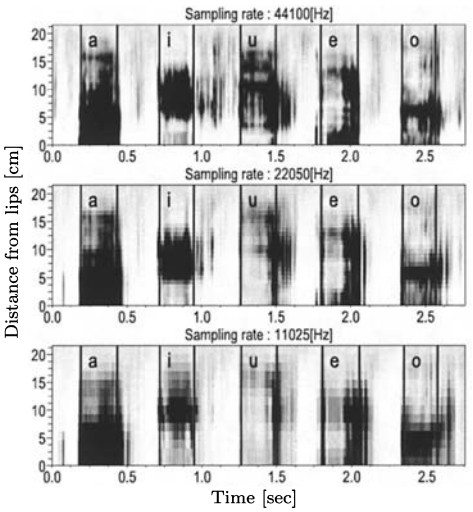

図 8　声道断面積関数の時間変化のサンプリング周波数による比較．上：サンプリング周波数 44100[Hz]，中：サンプリング周波数：22050[Hz]，下：サンプリング周波数：11025[Hz]

5.　話者変換への応用に関する検討

提案法の応用として，ある話者 A の音声を別の話者 B が話した音声のように声質制御することが可能であるか検討を行った．その方法として，/a/,/i/,/u/,/e/,/o/の 5 母音について，ある話者 A の各音素の声道断面積関数の平均値を，別の話者 B の各音素の声道断面積関数の平均値と同じ値になるよう声質制御を行った．これは，話者 A の元の声道断面積関数を $A_n^A(t)$，声質制御後の声道断面積関数を $\hat{A}_n^A(t)$，話者 A の声道断面積関数の平均値を μ_n^A，話者 B の声道断面積関数の平均値を μ_n^B とすると式 (10) で表される．T はフレーム数を表しており各音素により異なる．

$$\hat{A}_n^A(t) = A_n^A(t)\frac{\mu_n^A}{\mu_n^B} \quad t = 1, \cdots, T \tag{10}$$

図 9 に上記の声質制御により変化した共振周波数を示す．この値は式 (5) により得られたものである．声質制御に用いた音声はどちらも男性話者が「あいうえお」と発声したもので，サンプリング周波数は 11025Hz である．図の左は，ターゲットの話者の各音素の共振周波数を示しており，右は声質制御を行う話者の原音声の共振周波数と声質制御後の共振周波数を示している．図 10 に声道断面積関数ではなく PARCOR 係数で同様の処理を行った場合の共振周波数の変化を示す．図 9 と図 10 を比較すると提案法のほうがターゲットの共振周波数に近いことが確認できる．更に，図 11 に女性話者から男性話者へ声質制御した場合の共振周波数の変化を示す．この場合でも，声質制御後の声道断面積関数から求めた共振周波数はターゲットのものに近いことが確認できる．従って，提案法は性別に関係なく利用できるといえる．これらのことから，提案法は話者変換への応用に適していると考えられる．

6. ま と め

声道断面積関数による声質制御方式を提案した．

提案法に用いる音源と放射の特性の除去方法についての検討を行った．声質制御後の合成音声を試聴した結果，提案法では適応逆フィルタによりこれらの特性の除去を行うこととした．

声道断面積関数の形状はサンプリング周波数が異なる場合でも破綻せずに推定できることが分かった．このことは，提案法による声質制御が対象音声のサンプリング周波数に関係なく全て同じ枠組みで行うことができる可能性を示している．

また，話者変換への応用に関する検討を行った結果，5 母音に関してはターゲットに近づいたことを確認することができた．このことから，提案法は，話者変換への応用に適している考えられる．

文　献

[1] 中島隆之 他：“デコンボルーションによる声道形の推定と適応型音声分析システム”，日本音響学会誌，Vol.34, No.3, pp.157-166, 1978.

[2] Wakita,H. :“Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms”, IEEE Trans. Audio , Electroacoust., AU-21 , 5 , pp.417-427,1973.

[3] 河原英紀：“聴覚の情景分析が生んだ高品質 VOCODER : STRAIGHT”，日本音響学会誌，Vol.54, No.7, pp.521-526, 1998.

[4] 坂野秀樹 他：“包絡と音源の独立操作による音声モーフィング”，電子情報通信学会論文誌 A，Vol.J81-A，No.2, pp.261-268, 1998.

[5] 板倉，斉藤：“統計的手法による音声スペクトル密度とホルマント周波数の推定”，電子通信学会論文誌，53-A，1, pp.35-42, 1970.

[6] 太田，淵：“逆対象声道関数と母音特徴抽出”，日本音響学会誌，Vol.39，No.3，pp.173-184，1983.

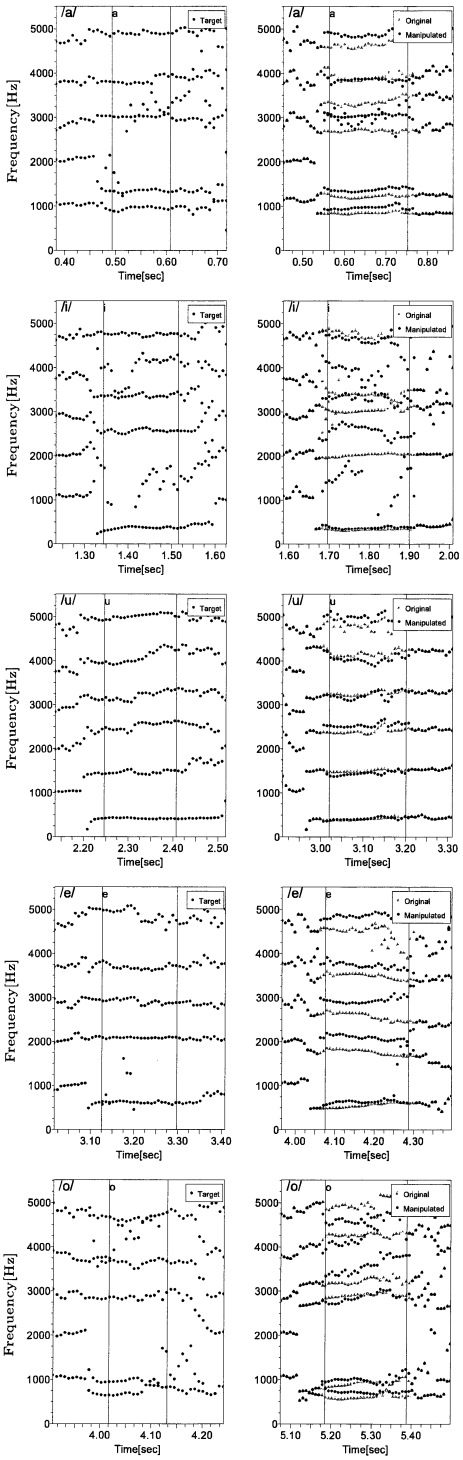

図 9　声質制御による共振周波数の変化．左：ターゲットとする話者の各音素の共振周波数．右：元の音声の共振周波数と声質制御後の共振周波数．

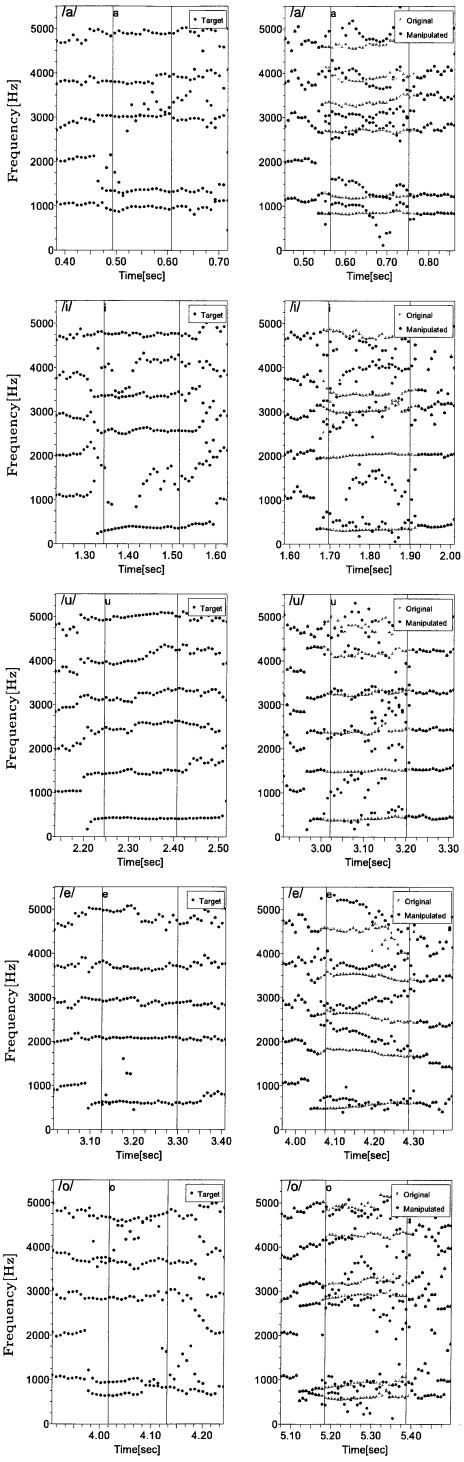

図 10　PARCOR 係数での声質制御による共振周波数の変化．左：ターゲットとする話者の各音素の共振周波数．右：元の音声の共振周波数と声質制御後の共振周波数．

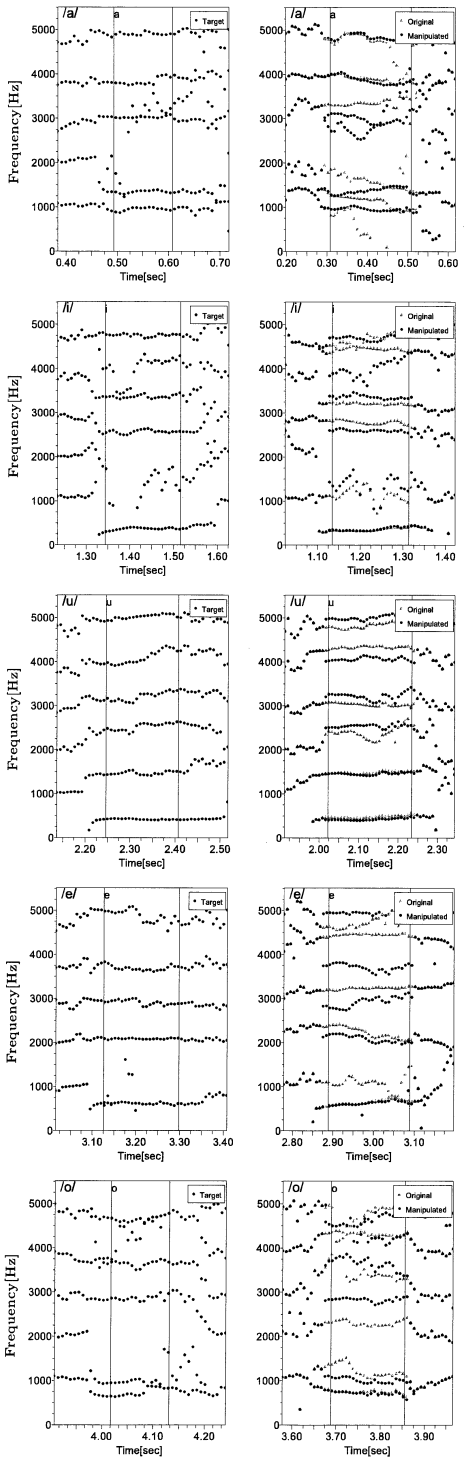

図 11　女性話者から男性話者への声質制御による共振周波数の変化．左：ターゲットとする話者の各音素の共振周波数．右：元の音声の共振周波数と声質制御後の共振周波数．