Vol.2020-MUS-127 No.17
Vol.2020-SLP-132 No.17
2020/6/6

複合正弦波モデルと動的計画法を用いた喉頭挙上音声への声質変換システムの提案

金井 郁也[1,a)]　荒川 薫[1,b)]　森勢 将雅[1,c)]

概要：2000 年代後半以降の DAW を用いたカラオケ音源の制作と流通により，既存の楽曲に自身が歌った音源を重ねた作品（歌ってみた）がウェブ上で多く投稿されている．それらの作品では質を高めるために，歌声に対して様々な補正が施されている．しかし，基本周波数やリズムの補正を行う歌唱補正システムはすでに提案されているが，声質に関する補正を行うシステムは不足している．本稿では，音声における音色変化のうち，喉頭の高さの変化が音色に与える影響の調査をし，喉頭挙上によるスペクトル包絡の変化を考慮した声質変換システムの提案と本システムの評価実験を行った．実験の結果，通常の状態で発声した音声を分析後そのまま再合成した音声と比較して，基本周波数を上昇させたうえで本システムの処理を行った音声の喉頭の高さに相当する評価結果が有意に上昇した（$p < 0.01$）．

1. はじめに

2000 年以降のパソコンの普及に伴いアマチュアを含めた多くの人が楽曲制作や録音を容易に行える環境にあり，ディジタル・オーディオ・ワークステーション（Digital Audio Workstation : DAW）を用いたコンテンツを作成している．その中で近年注目されているものの 1 つが「歌ってみた」である．

作曲者本人がカラオケ音源を公開している場合が多い VOCALOID [1] を用いた楽曲が動画投稿サイトに投稿されるようになってから，カラオケ音源を入手することは容易になった．DAW を用いて作成した既存の楽曲のカラオケ音源を無料公開している人も少なくない．こうした近年のカラオケ音源の提供・入手環境の変化に伴い，多くの人が「歌ってみた」の動画や音源を作成・共有出来るようになった．しかし，作品の質を高めるには様々なエフェクターを扱える必要があり，その技術を身に着けるには多くの時間を要する．新規のユーザーが質の高い作品を作るためには，音源を自動で修正するシステムや少ないパラメータの操作で意図した通りの変換を行えるエフェクターが必要である．

基本周波数やリズムの補正を行う歌唱補正システムはすでに提案されているが，声質に関する補正を行うシステムが不足している．先行研究で提案されている声質も含めた歌唱補正システム [2], [3] は特定の人物の音声データを用いて学習させた声質変換であり，学習データが不足している場合への対応がなされていない．声質変換を行えるエフェクターも存在するが，変換目標とする音声が存在しない声質変換は不自然な音声になりやすい．以上のような問題は，声道の形状の変化が定式化されていないことが一因になっていると考えられる．人の声道とは，一般に発声器官のうち声帯より上に位置する喉頭・咽頭・鼻腔・口腔を指す（図 1）．声帯の振動で生じた音に声道形状による周波数特性（スペクトル包絡）を与えることで，人は様々な声を発することが出来る．そのため，ユーザーが意図したとおりの声質に変換するためには，ユーザーが想像する音声の特徴とスペクトル包絡の対応関係を調査する必要がある．

本研究では，音声における音色変化のうち，喉頭の上昇（喉頭挙上）に伴う音色変化を制御する声質変換システムを提案する．また，本システムの主観評価実験を行った結果，喉頭の高さに関して，通常の状態で発声した音声を分析後そのまま再合成した音声と，基本周波数を上昇させたうえで本システムの処理を行った音声の間には有意差が認められた．

2. 喉頭の高さの変化が音声に与える影響

喉頭の高さの変化が音声に与える影響の 1 つがフォルマントのシフトであると先行研究 [4], [5] で述べられている．フォルマントとは音声の周波数スペクトルの包絡線（スペクトル包絡）における複数のピークを指す．喉頭の高さの変化によるフォルマントのシフト（図 2）は，共鳴管（声

[1] 明治大学大学院先端数理科学研究科
[a)] cs202003@meiji.ac.jp
[b)] kara@meiji.ac.jp
[c)] mmorise@meiji.ac.jp

Vol.2020-MUS-127 No.17
Vol.2020-SLP-132 No.17
2020/6/6

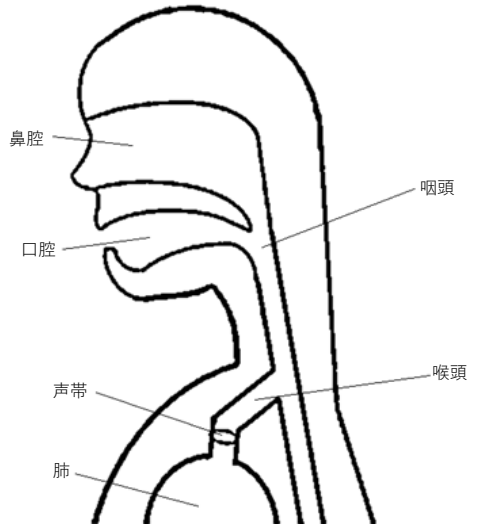

図 1 発声器官の構造

道）の長さの変化とともに共鳴周波数も変化するために生じると考えられる．また，喉頭の高さの変化は声道の長さだけでなく声道（咽頭）の形状にも影響を与えるとも述べられている．

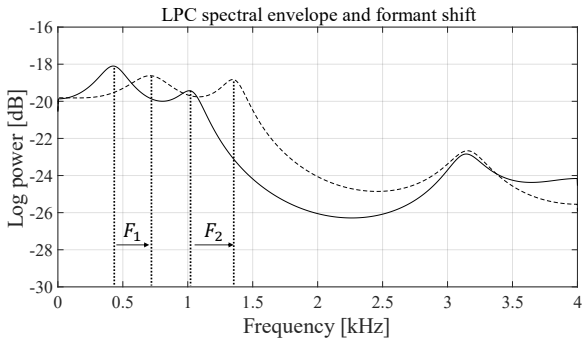

図 2 通常音声の LPC スペクトル包絡（実線）と喉頭挙上音声の LPC スペクトル包絡（破線）とフォルマント周波数（点線）

3. 喉頭挙上に相当する音色変化を与えるシステムの実装

本研究では，少ないデータに基づく声質変換システムの実現を目的とするため，声質に起因する特徴量であるスペクトル包絡に対してのみ変換した．本システムは参照音声から特徴量を抽出する抽出領域と入力音声を変換する変換領域の 2 つで構成されている．システム構成を図 3 に示す．

3.1 特徴量の抽出

抽出領域では喉頭挙上によって起きるスペクトル包絡の変化を実装するための特徴量を抽出する．その変化の 1 つがフォルマントのシフトだが，フォルマントの抽出は不安定で声質変換に用いるのは難しい．本システムでは坂野らの研究 [6] を参考に，通常の喉頭の位置で発声した音声（通常音声）と喉頭を上げた状態で発声した音声（喉頭挙上音声）のスペクトル包絡の周波数軸上での動的計画法（Dynamic Programming : DP）マッチングにより求まるワーピングパス（歪み関数）を用いてスペクトル包絡の非線形伸縮を行うことでフォルマントのシフトを実装した．また，先行研究 [4] から母音によって変換式が異なることが想定されるため，全ての特徴量を母音ごとに抽出し，それらを「通常音声の各母音」と「入力音声の各フレームにおける母音」の距離を用いて重みづけする．この距離を求めるのに有効な特徴量の 1 つがフォルマントだが，先述の通りフォルマントの安定した抽出は困難であるため，本研究では嵯峨山らが提案した複合正弦波モデル（Composite Sinusoidal Model : CSM）[7] によって抽出される CSM 周波数を用いる．CSM 周波数はフォルマントに類似した特徴量でかつ安定した抽出が可能な特徴量であり，これを用いることで入力音声の各フレームの母音によって異なる変換式を得る．また，スペクトル包絡の値は対数パワー値で表されるものとする．

まず，複合正弦波モデルによって通常音声の 1 次，2 次 CSM 周波数を抽出し，それぞれの母音ごとの平均を求める．次に，母音 v・音名 l で発声した「通常音声の i フレーム目におけるスペクトル包絡」と「喉頭挙上音声の i フレーム目におけるスペクトル包絡」の周波数軸上での DP マッチングによりワーピングパス $p_{v,l}(i)$ を求める．ここで v は母音の集合 V の要素を表す変数とし，l は実験で用いたオクターブを考慮した音名の集合の要素 L を表す変数とする．オクターブを考慮するため，例えば A3 と A4 は異なる要素となる．通常音声と喉頭挙上音声のスペクトル包絡の DP マッチングによって生成されたワーピングパスの例を図 4 に示す．このワーピングパスの母音ごとの平均を P_v とする（式 (1)）．また，計算量を減らすために DP マッチングは一定のフレーム数 N の間でのみ行う．

$$P_v = \frac{1}{N|L|} \sum_{k \in L} \sum_{j=1}^{N} p_{v,k}(j). \tag{1}$$

各母音，音名，フレームで「喉頭挙上音声のスペクトル包絡 $\mathrm{HS}_{v,l}(i)$」と「通常音声のスペクトル包絡 $\mathrm{NS}_{v,l}(i)$ をワーピングパス $p_{v,l}(i)$ を用いて伸縮を行ったスペクトル包絡」の差分を母音ごとに平均化したものを伝達関数 f_v とする（式 (2)）．これを最小二乗法によって n 次多項式に近似した伝達関数 F_v を求める（式 (3)）．f_v と F_v の例を図 5 に示す．

$$f_v = \frac{1}{N|L|} \sum_{k \in L} \sum_{j=1}^{N} \Big(\mathrm{HS}_{v,k}(j) - \mathrm{Shift}\big(\mathrm{NS}_{v,k}(j), p_{v,k}(j)\big) \Big), \tag{2}$$

$$F_v = \sum_{k=0}^{n} a_{v,k} x^k \simeq f_v, \tag{3}$$

ここで Shift は与えられたワーピングパスに基づいてスペ

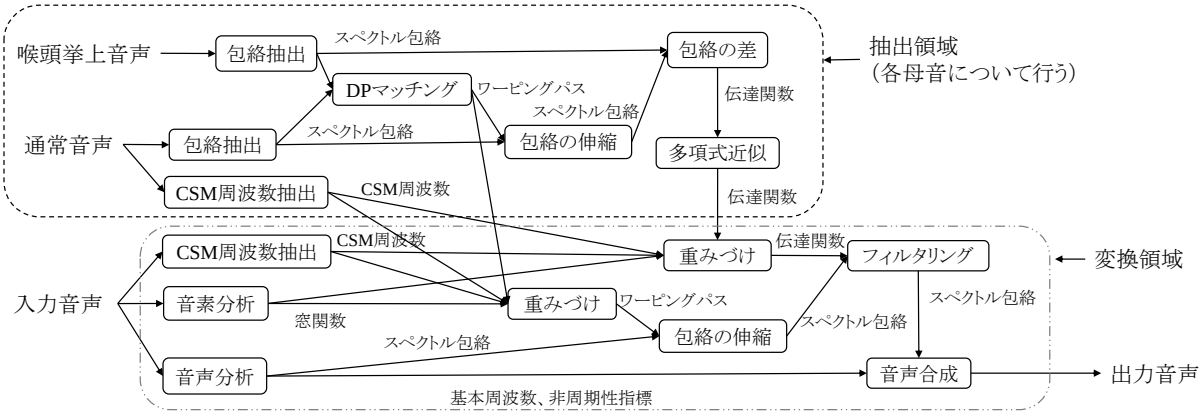

図 3 システム構成

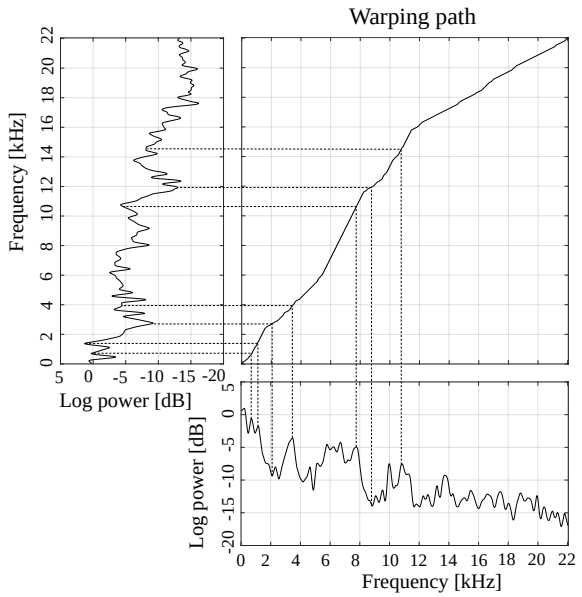

図 4 通常音声のスペクトル包絡（右下）と喉頭挙上音声のスペクトル包絡（左上）とワーピングパス（右上）

クトル包絡の伸縮を行う関数である．

データが複数人分ある場合，変換領域で用いる特徴量である各母音の平均の 1 次，2 次 CSM 周波数，ワーピングパス P_v，伝達関数 F_v は全員の値の平均値とする．

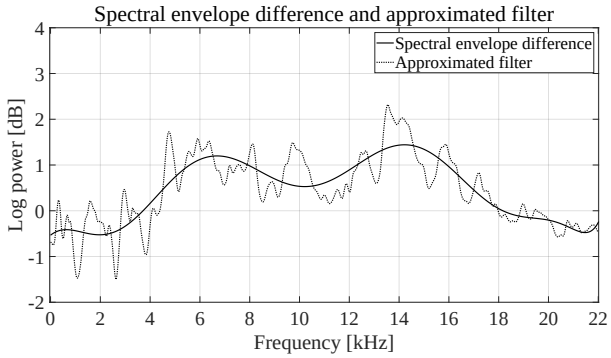

図 5 平均化した伝達関数（点線）と
10 次多項式で近似した伝達関数（実線）

3.2 合成音声の生成

変換領域では，抽出領域で求めた特徴量を用いて入力音声の声質変換を行う．まず，i フレーム目における入力音声の 1 次，2 次 CSM 周波数と抽出領域で得た母音 v を発声した通常音声の 1 次，2 次 CSM 周波数のユークリッド距離 $D_v(i)$ の逆数を正規化した値 $I_v(i)$ を求める（式 (4)）．

$$I_v(i) = \frac{1}{D_v(i)\sum_{k\in V}\frac{1}{D_k(i)}}. \tag{4}$$

次に，入力音声の i フレーム目における音素 $\mathrm{Phone}(i)$ を音素分析により推定し，i フレーム目が無声子音の時のみ重み係数を 0 にする窓関数 $w(i)$ を設計する（式 (5)）．

$$w(i) = \begin{cases} 0 & \big(\mathrm{Phone}(i) \in U\big) \\ 1 & \big(\mathrm{Phone}(i) \notin U\big) \end{cases}, \tag{5}$$

ここで U は全ての無声子音の集合とする．図 6 に音声の波形と音素から導出された窓関数の例を示す．

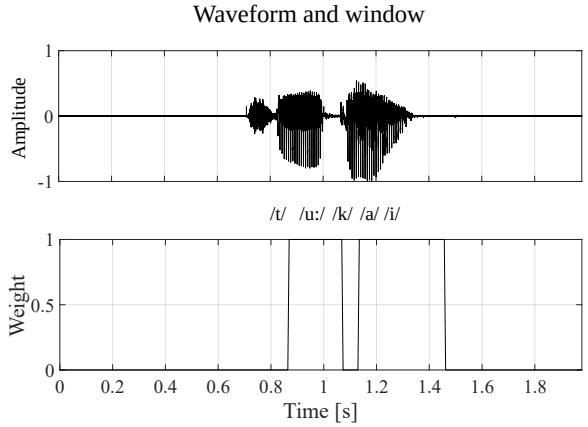

図 6 音声の波形（上）と音素から導出された窓関数（下）

$I_v(i)$ に窓関数 $w(i)$ をかけることにより，i フレーム目における P_v，F_v に対する重み係数 $W_v(i)$ が求まる（式 (6)）．

$$W_v(i) = w(i)I_v(i). \tag{6}$$

抽出領域で求めた P_v，F_v を用いて重みづけを行うことで，i フレーム目におけるワーピングパス Path(i)，伝達関数 Filter(i) を求める（式（7），式（8））．

$$\mathrm{Path}(i) = \mathrm{NP} + \sum_{k \in V} (P_k - \mathrm{NP}) W_k(i), \tag{7}$$

$$\mathrm{Filter}(i) = \sum_{k \in V} F_k W_k(i), \tag{8}$$

ここで NP は DP マッチングにおける傾き 1 のワーピングパスとなる．

ワーピングパス Path(i) を用いて入力音声の i フレーム目におけるスペクトル包絡 IS(i) の伸縮を行い，伝達関数 Filter(i) をフィルタとして用いることで変換後のスペクトル包絡 OS(i) が求まる（式（9））．

$$\mathrm{OS}(i) = \mathrm{Shift}\big(\mathrm{IS}(i), \mathrm{Path}(i)\big) + \mathrm{Filter}(i). \tag{9}$$

式（9）を入力音声の全フレームに適用することにより合成音声が生成される．

4. 参照音声の録音と分析

本実験では，本システムの声質変換に必要な特徴量を被験者の男性 2 名から抽出した．被験者には，通常音声と喉頭を上げた状態の音声を 1 オクターブの音域で日本語における 5 母音を半音ごとに発声するよう指示を行った．また，スペクトル包絡抽出には森勢が開発した CheapTrick [8] を用いた．

表 1 録音・分析条件

録音場所	防音室（D-40）
録音日時	2019 年 9 ～ 12 月
サンプリングレート	44.1 kHz
使用音域	C3（130.8 Hz） ～ C4（261.6 Hz）
量子化ビット数	16 bit
フレーム長	2048 サンプル（46 ms）
フレームシフト	5 ms
DP マッチングを行ったフレーム数	100 frame（500 ms）
DP マッチングの傾斜制限	0.5 ～ 2
フィルタの n 次多項式近似	$n = 10$

録音した音声の分析を行い，抽出された 2 人の被験者の平均の CSM 周波数の分布図を図 7 に，各母音におけるワーピングパスから求めたスペクトル包絡の各周波数における伸縮倍率を図 8 に，各母音における伝達関数を図 9 に示す．また，10 次多項式で近似されている伝達関数の各項の値は表 2 のようになった．

5. 合成音声の主観評価実験

本実験では入力音声と比較した際の合成音声の喉頭の高さ，個人性，音質について調査するための主観評価実験を行った．

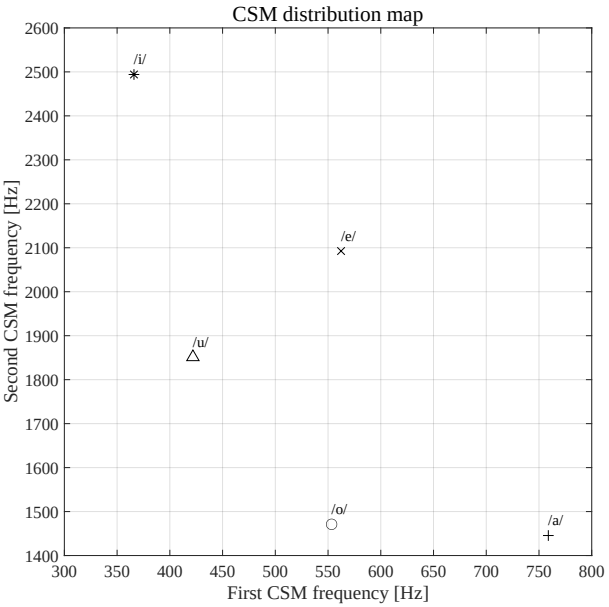

図 7 被験者の平均の CSM 周波数分布

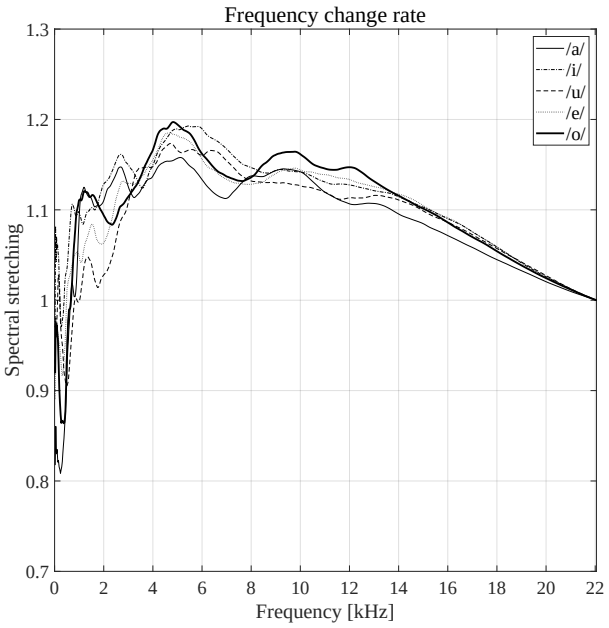

図 8 母音ごとのスペクトル包絡の各周波数における伸縮率

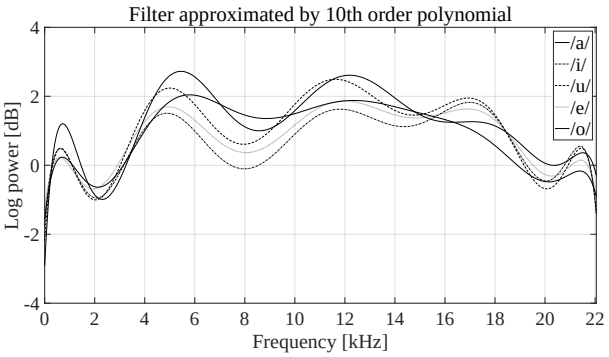

図 9 各母音における伝達関数

表 2　各母音における伝達関数の項の値

	/a/	/i/	/u/	/e/	/o/
a_{10}	$-6.29E-39$	$-1.45E-38$	$-1.65E-38$	$-1.05E-38$	$-1.26E-38$
a_9	$7.19E-34$	$1.62E-33$	$1.85E-33$	$1.17E-33$	$1.46E-33$
a_8	$-3.53E-29$	$-7.67E-29$	$-8.78E-29$	$-5.61E-29$	$-7.25E-29$
a_7	$9.67E-25$	$2.02E-24$	$2.32E-24$	$1.49E-24$	$2.01E-24$
a_6	$-1.62E-20$	$-3.23E-20$	$-3.72E-20$	$-2.40E-20$	$-3.39E-20$
a_5	$1.70E-16$	$3.22E-16$	$3.71E-16$	$2.39E-16$	$3.59E-16$
a_4	$-1.10E-12$	$-1.96E-12$	$-2.26E-12$	$-1.45E-12$	$-2.34E-12$
a_3	$4.16E-09$	$6.87E-09$	$7.97E-09$	$5.07E-09$	$8.88E-09$
a_2	$-8.07E-06$	$-1.24E-05$	$-1.44E-05$	$-8.93E-06$	$-1.76E-05$
a_1	$6.52E-03$	$9.07E-03$	$1.07E-02$	$6.37E-03$	$1.45E-02$
a_0	$-1.57E+00$	$-1.76E+00$	$-2.19E+00$	$-1.30E+00$	$-2.92E+00$

5.1　提示する音声の生成

1 人の男性の被験者に通常の状態で 10 種類の単語を発声してもらった音声を元に合成音声の生成を行った．音声の収録は 2020 年 1 月に行い，録音場所，サンプリングレート，量子化ビット数，フレーム長，フレームシフトは表 1 と同様の条件で行った．また，CSM 周波数抽出時のダウンサンプリングも同様である．

実験に用いた 10 種類の単語を表 3 に示す．1 つの単語につき，分析後そのまま再合成した音声，音高を全音上げた音声，本システムによってフォルマントのシフトをした音声，音高を全音上げたうえで本システムによってフォルマントのシフトをした音声の 4 種類の合成音声を生成した．一般に発声周波数と喉頭の高さには相関関係が認められている [9] ため，音高の変化を加えた合成音声も生成した．入力音声，2 秒間の無音，合成音声，3 秒間の無音の順に流れる音源を単語の種類分（10 回）つなぎ合わせたデータを作成する．このデータを音声群と呼ぶ．提示する合成音声は全部で 4 種類であるため，合計で 4 つの音声群が作成されることになる．

表 3　実験に用いた単語

単語	読み
痛快	つーかい
自動ドア	じどーどあ
百億	ひゃくおく
税金	ぜーきん
リアリティ	りありてぃ
小学校	しょーがっこー
一般論	いっぱんろん
現代	げんだい
ファミリー	ふぁみりー
目上	めうえ

5.2　実験手順

前節で生成した合成音声を用いて，9 人の被験者による主観評価実験を行った．はじめに，被験者に喉頭が上がった音声と下がった音声の例を聴いてもらい，それぞれの音声に関する簡潔な説明をする．次に音声群のうちの 1 つを提示し，入力音声と比較した際の合成音声の喉頭の高さ，個人性，音質について 7 段階で評価，すなわち，喉頭の高さが「非常に下がっている」と感じたら 1,「非常に上がっている」と感じたら 7,「変わっていない」と感じたら 4 と，個人性が「全く保たれていない」と感じたら 1,「完全に保たれている」と感じたら 7,「どちらとも言えない」と感じたら 4 と，音質が「非常に悪くなった」と感じたら 1,「非常に良くなった」と感じたら 7,「変わっていない」と感じたら 4 と評価してもらう．この評価を全ての音声群に対して行った後，実験を終える．4 種類の音声群を提示する順番は被験者によってランダムに変更した．

基本周波数抽出には雑音に強い Harvest [10] を，スペクトル包絡抽出には CheapTrick [8] を，非周期性指標抽出には D4C [11] を，音声合成には World [12] を用いた．また，音素分析には Julius [13] の音素セグメンテーションキットを用いた．

5.3　実験結果と考察

実験の結果を表したグラフを図 10 に示す．ここで「分析後そのまま再合成した音声」を None,「音高を全音上げた音声」を Pitch,「本システムの処理を行った音声」を Shift,「音高を全音上げたうえで本システムの処理を行った音声」を Both とする．分散分析の結果，喉頭の高さ

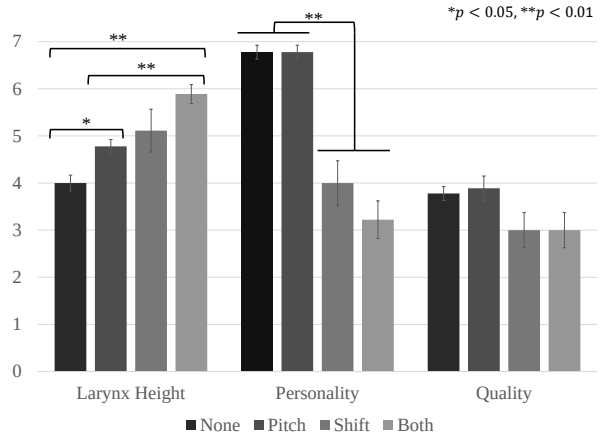

図 10　主観評価実験による喉頭の高さ，
個人性，音質の評価の平均値と標準偏差

（Larynx Height）と個人性（Personality）にはそれぞれ有意差（$p < 0.01$）が認められたため，喉頭の高さと個人性については Steel-Dwass 法を用いた多重比較検定も行った．

喉頭の高さについての多重比較検定の結果，None-Pitch，None-Both，Pitch-Both 間で有意差が認められた（None-Pitch : $p < 0.05$，None-Both・Pitch-Both : $p < 0.01$）．しかし，None-Shift 間で有意差は認められなかったため，本システムの処理だけでは喉頭挙上による音色変化が生じるとは言えない．また，基本周波数の増加を行った方が評価が高くなることが分かった．これは，喉頭が上がった音声における高周波数帯域のブーストと基本周波数の増加を混同するためだと推測される．

個人性についての多重比較検定の結果，None-Shift，None-Both，Pitch-Shift，Pitch-Both 間で有意差が認められた（$p < 0.01$）．このことから，本システムに処理された音声の個人性は低下すると考えられる．しかし，実際の音声も喉頭を上げることによって個人性が下がる可能性があるため，合成音声ではない通常音声と喉頭挙上音声の個人性に関する主観評価実験により，個人性の低下が実際に起きる現象かを調査する必要がある．

6. おわりに

本研究では，喉頭の高さが音声に与える影響の調査と喉頭挙上がスペクトルに与える影響を考慮した声質変換システムを提案した．調査の結果，喉頭の高さが変わることで声道の長さと形状が変化するということが分かった．その中でも声道の長さの変化に伴うフォルマントの変化に着目し，DP マッチングによってフォルマントのシフトの実装を行うシステムを構築した．

主観評価実験と分析の結果，喉頭の高さに関して，通常の状態で発声した音声を分析後そのまま再合成した音声と本システムで処理された音声の間には有意差は認められなかった．しかし，基本周波数を上昇させたうえで本システムの処理を行った音声との間には有意差が認められたため，音声から喉頭の上昇を感じるにはフォルマントのシフトと基本周波数の上昇が必要であると考えられる．同時に本システムで処理された音声には個人性の低下が生じることも明らかになった．

今後は個人性の大幅な低下が実際の通常音声と喉頭挙上音声の間にも起きる現象かを調査するとともに，より多くの音声データを収集することでシステムの質の向上を図る．

参考文献

[1] 剣持秀紀，大下隼人ほか：歌声合成システム VOCALOID-現状と課題，情報処理学会研究報告音楽情報科学 (MUS)，Vol. 2008, No. 12 (2008-MUS-074), pp. 51–56 (2008).
[2] 中野皓太，森勢将雅，西浦敬信，山下洋一：基本周波数の転写に基づく実時間歌唱制御システムの実現を目的とした高品質ボコーダ STRAIGHT の高速化，電子情報通信学会論文誌 A， Vol. 95, No. 7, pp. 563–572 (2012).
[3] Nakano, T. and Goto, M.: VocaListener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics, *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 453–456 (2011).
[4] Sundberg, J. and Nordström, P.-E.: Raised and lowered larynx-the effect on vowel formant frequencies, *STL-QPSR*, Vol. 17, No. 2-3, pp. 035–039 (1976).
[5] Sundberg, J. and Askenfelt, A.: Larynx height and voice source: a relationship?, *STL-QPSR*, Vol. 22, No. 2-3, pp. 023–036 (1981).
[6] 坂野秀樹，武田一哉，鹿野清宏，板倉文忠：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. 81, No. 2, pp. 261–268 (1998).
[7] 嵯峨山茂樹，板倉文忠：複合正弦波モデルによる音声スペクトルの分析，電子情報通信学会論文誌 A， Vol. 64, No. 2, pp. 105–112 (1981).
[8] Morise, M.: CheapTrick, a spectral envelope estimator for high-quality speech synthesis, *Speech Communication*, Vol. 67, pp. 1–7 (2015).
[9] Pabst, F. and Sundberg, J.: Tracking multi-channel electroglottograph measurement of larynx height in singers, *STL-QPSR*, Vol. 33, No. 2-3, pp. 067–078 (1992).
[10] Morise, M. et al.: Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals., *INTERSPEECH*, pp. 2321–2325 (2017).
[11] Morise, M.: D4C, a band-aperiodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol. 84, pp. 57–65 (2016).
[12] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884 (2016).
[13] 河原達也，李晃伸：連続音声認識ソフトウェア Julius(＜特集＞研究のツールボックス (2))，人工知能学会誌，Vol. 20, No. 1, pp. 41–49 (2005).