電気式人工喉頭を用いた歌唱システムにおける自然な身体動作を利用した歌唱表現付与の提案

大川 舜平[1,a)]　石黒 祥生[2,b)]　大谷 健登[1,c)]　西野 隆典[4,d)]　小林 和弘[3]　戸田 智基[3,e)]
武田 一哉[2,f)]

概要：喉頭がんや咽頭がんなどの治療のため喉頭摘出手術を受けた人々（喉頭摘出者）は自身の声帯を用いた発声を行うことができず，代替発声法として電気式人工喉頭（Electrolarynx:EL）を使用することが多い．しかし，EL による発話音声は機械的であり，さらに音高の制御が困難であるため，歌を歌いたいという需要に答えることができない．そこで従来研究では，EL による発話音声を自然な声質へと変換する処理により得られた特徴量系列と，MIDI 音源を人間らしい音高推移へと変換する処理により生成された音高パターンを合成することにより自然な歌声へと変換するシステムを実現した．一方で，この手法は歌唱中，ユーザの意思通りに歌唱表現を制御することができないという課題がある．この問題に対して本研究では歌唱中に生じる自然な身体表現を用いることで歌唱表現を付与することが可能であるかを検討する．本稿では顔のランドマーク推定技術や全身の骨格推定技術を用い，「腕を伸ばす」等の特定の動作を認識することでビブラートの強さを制御することが可能なシステムを提案した．結果，表情による制御は繊細な調整が困難であり，意図しない箇所にビブラートが付与された一方で，腕による操作では表現性及び操作性において肯定的な意見が得られた．また，提案システムの処理遅延は約 160 ms であったが，被験者の多くは遅延を感じなかったと回答した．

1. はじめに

発声や歌唱は我々が持つ基本的なコミュニケーション手段のひとつであり，自己を表現する上で重要な役割を担う．一方で，喉頭がんや咽頭がんなどの治療のため喉頭摘出手術を受けた人々（喉頭摘出者）は自身の声帯を用いた発声を行うことができず，何らかの代替発声手段が必要となる．

喉頭摘出者の代表的な代替発声法のひとつに電気式人工喉頭（Electrolarynx:EL）を用いた発声法が挙げられる．EL は，喉付近に当て，音源振動を体外から声道に送ることで擬似的な声帯の役割を果たし，これにより喉頭摘出者は通常の発話と同様に口を動かすことで電気音声として発声を行うことができる．さらに，EL には歌唱を行う機能も搭載されている（ユアトーン，株式会社電制）．しかし，この機能はボタンを押すことで音高を変化させるという仕組みであり，音高を変化させる度に音が途切れてしまうため，実際の歌唱と比較して不自然である．

これに対し，森川らは統計的声質変換による電気音声の声質変換と MIDI キーボードによる基本周波数（F_o）制御を組み合わせることで，人間らしい声でありながらメロディを自由に操作可能なシステムを提案した [1]．これにより喉頭摘出者は MIDI キーボードを用いて自由な歌唱を行うことが可能になった．また，[1] ではルールベースのフィルタリングを用いることで音高処理を行っているが，現在 VAE を用いることで F_o 制御性能を改善するための取り組みがなされている [2]．

一方で，実際の歌唱と比較すると，森川らによる手法には表現性において不十分な点がいくつかある．ひとつは MIDI キーボード入力による音高制御では「ビブラート」や「こぶし」，「しゃくり」のような繊細な音高制御を行うことができないことである．また，声質変換において，あらかじめ学習した声質に依存しているため，「グロウル」や「ウィスパー」といった声質を切り替える操作は困難である．

そこで，本研究では歌唱表現の際に伴う特定の手足の動作や表情の変化などの自然な身体表現に着目し，これらを

[1] 名古屋大学 情報学研究科
[2] 名古屋大学 未来社会創造機構
[3] 名古屋大学 情報基盤センター
[4] 名城大学 都市情報学部
[a)] okawa.shumpei@g.sp.m.is.nagoya-u.ac.jp
[b)] ishiy@acm.org
[c)] ohtani.kento@g.sp.m.is.nagoya-u.ac.jp
[d)] nishino@meijo-u.ac.jp
[e)] tomoki@icts.nagoya-u.ac.jp
[f)] kazuya.takeda@nagoya-u.ac.jp

電気音声変換技術における制御パラメータとしてリアルタイムに反映することで歌い手の意思に応じた歌唱表現の付与を実現する手法の検討を行う．

2. 関連研究

2.1 喉頭摘出者のための発話・歌唱支援

喉頭摘出者が自然な音声での発話を可能にする手法の 1 つに統計的声質変換 [3] がある．これは機械学習を用いることで EL による発声を人間の声に変換するという技術であり，この音声に対して雑音除去 [4] や音質改善 [5] を行うことで，より質の高い音声変換を実現しようとする取り組みが行われている．これに対して，雑音が多いという EL の欠点を克服するために他の音源を録音することで発話支援を行う研究も存在する．例えば，中村ら [6] は微弱な外部音源信号を用いて発声された音声を体表密着型のマイクロホンを用いて収録し，統計的声質変換を用いることで，より自然なささやき声の生成を実現した．

また，前述した発話音声に対して F_o 制御を行うことでより自然な音高変化を持った音声を生成することができる．田中ら [7] は電気音声から F_o パターンをリアルタイムで予測する統計的音源予測処理を提案し，音声の自然性を大幅に改善することができる可能性を示した．

一方で，森川ら [1], [2] は MIDI キーボードにより F_o 制御を行うことで歌唱することができるシステムを提案した．このシステムは主に 2 つの制御により構成されている．ひとつは電気音声から肉声への変換処理部であり，CLDNN による声質制御を行うことで電気音声を肉声に音声特徴量系列を変換することができる．もう一方は MIDI 音高パターンを人間の歌声に近い F_o 変動パターンに変換するというもので，これに前者によって得られた音声特徴量系列を合成することで歌声の生成を可能にする．本研究ではこのシステムに対して，歌唱者の意思に応じて歌唱表現を付与することが可能にするためリアルタイムで F_o 及び声質を制御する手法を検討する．

2.2 身体表現の認識技術

歌唱表現の際に伴う特定の手足の動作や表情の変化を認識し，それを歌声加工の起点とするためにはリアルタイムでの表情やポーズの認識技術が不可欠である．表情認識技術においては，リアルタイムで顔のランドマーク推定を行うことができる技術は数多く存在しており [8], [9]，さらにランドマーク情報から表情や感情を推定する技術 [10] も提案されている．

一方で，リアルタイムでジェスチャーを認識する技術として Cao ら [11] は OpenPose を提案した．OpenPose は 2D の映像からリアルタイムで複数の人数に対してジェスチャー認識可能な技術であり，骨格や指，顔のパーツなど計 135 のポイントを認識することができる．さらに，Schneider ら [12] は OpenPose を使用して人物のポーズを抽出する RGB ビデオでのポーズ推定と，時系列分類用の One-Nearest-Neighbor と組み合わせた動的タイムワーピングを提案し，「腕を組む」や「右腕をスワイプ」などの計 6 つのポーズの判別を実現した．

これらの技術を用いることで例えばビブラートの場合，「口を開く大きさを素早く変化させる」，こぶしの場合「こぶしを掲げる」というようなジェスチャーを認識することで表現付与の起点とすることが可能であると考えられる．

3. 提案手法

3.1 システムの概要

本稿では森川ら [1], [2] による歌声生成システムにおける音高パターン制御部において，身体表現により得られるパラメータ情報を統合することでリアルタイムで歌唱表現を付与可能なシステム（図 1）を提案する．

なお，森川らによる手法では歌唱することができる楽曲が限られてしまうという課題を解決するために，歌唱中に MIDI キーボードを演奏することで音高情報を入力しているが，本論文では身体表現を用いる上で妨げとなることを考慮し，あらかじめ用意した MIDI 音源を用いる．

3.2 音高制御手法

実際の歌唱における F_0 には次のような動的変動成分が含まれることが分かっている（図 2）．

オーバーシュート
滑らかな音高の変化，およびその直後に目的音高を越える瞬時的な変動成分

ビブラート
同一音高区間で観測される 4-8 Hz の準周期的な変動成分

プレパレーション
音高変化直前に変化とは逆方向に振れる瞬時的な変動成分

微細変動
発声区間全体に観測される不規則で細かい変動成分

森川らによる F_o 制御モデル [1], [2], では MIDI 入力に対して以下の伝達関数を適用することで，オーバーシュート，ビブラート，プレパレーションを含んだ F_o 推移を生成可能となる．

$$H(s) = \frac{k}{s^2 + 2\zeta\omega s + \omega^2} \tag{1}$$

ここで，ω は固有角周波数，ζ は減衰項，k は振幅項を表し，これに対応するインパルス応答は以下に示す通りである．

$$h(t) = \begin{cases} kt\exp(-\omega t) & 0 < |\zeta| < 1 \\ \frac{k}{\omega}\sin(\omega t) & |\zeta| = 0 \end{cases} \tag{2}$$

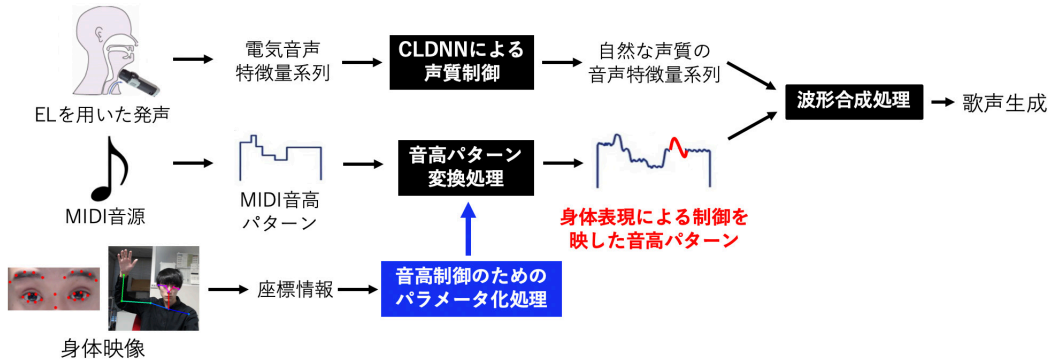

図 1　提案システムの概要

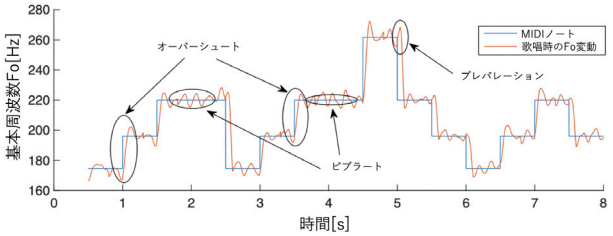

図 2　歌唱における F_o 推移と MIDI によるステップ状の推移

なお，$\lambda_1 = -\zeta + \sqrt{\zeta^2 - 1}$，$\lambda_2 = -\zeta - \sqrt{\zeta^2 - 1}$ であり，オーバーシュートとプレパレーションは減衰振動モデル（$0 < |\zeta| < 1$），ビブラートは定常振動モデル（$|\zeta| = 0$）で記述される．また，これらは現在の音高情報，直前の音高情報，直前の音高が継続している時間の 3 つの情報から選択される．具体的には音高変化時から 200 ms はオーバーシュートが選択され，さらに同一音高が続く場合はビブラートが選択される．

本稿では身体表現を用いたパラメータ α を用いることにより，式（3）のようにビブラート制御部における F_o 定常振動の振幅を制御する．

$$h(t) = \left(\frac{k}{\omega} + \alpha\right) \sin(\omega t) \tag{3}$$

3.3　身体表現を用いた制御手法

歌唱の際に現れる身体表現として，表情の変化，そして腕など全身の動作が挙げられる．そこで本研究では，これらの身体表現認識するために映像情報を利用し，表情の変化を顔のランドマーク推定技術，全身の動作を骨格推定技術を用いることで特定の動作の識別を行う．ただし，リアルタイムでの身体表現によるパラメータ制御には低遅延性が求められるため，特定の動作の識別には複雑な認識処理を使用せず，座標情報を用いた簡潔なシステムを検討する．

3.3.1　ポーズ推定技術を用いた腕による制御

提案システムには OpenPose[11] の tensorflow[13] 版である tf-pose-estimation を用いて 2 種類の特定の動作による制御手法を提案する．ひとつは「手の上げ下げ」による制御である．これは図 3 に示すように手首と肘の座標間

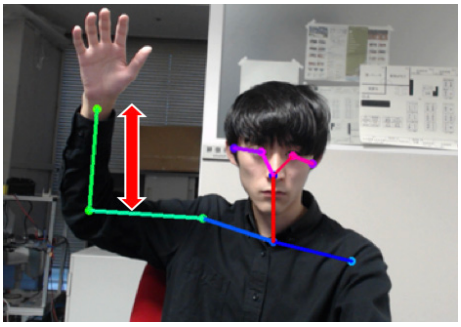

図 3　手の上げ下げによる制御

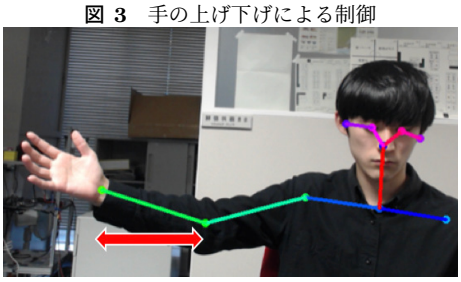

図 4　腕の伸ばし縮みによる制御

の鉛直方向の距離をパラメータとして用いるというものであり，手を上げる，つまり距離が大きくなるに連れビブラートの F_o 振幅パラメータは大きな値をとる．もう一方は「腕の伸ばし縮み」による制御（図 4）であり，こちらでは水平方向の距離を用いることで制御を行う．なお，tf-pose-estimation における 1 フレームあたりの処理時間は約 75 ms である．

3.3.2　顔のランドマーク検出を用いた眉毛による

顔のランドマーク検出には，Xuanyi ら [8] によって提案された Supervision by Registration を用いる．この手法では目，眉毛，唇など計 68 個の座標をリアルタイムで取得することが可能であり，1 フレームあたりの処理時間は約 60 ms である．本研究ではこの技術を用いることで「眉毛の上げ下げ」による制御を提案する．具体的には図 5 に示すように，（鼻の上部から眉毛からの y 座標距離）/（鼻の長さ）が閾値より大きな値である場合は「眉毛を挙げている」，閾値未満である場合は「眉毛を下げている」と認識し，これにより得られた値をパラメータとして用いる．また，

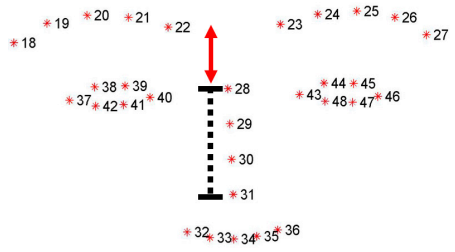

図 5 眉毛の上げ下げによる制御

この手法による認識は被験者間での閾値の個人差が大きいため，被験者ごとに閾値や変化の割合の調整を行った．

4. 実験

本論文ではビブラートの制御を行わない従来手法と比較してビブラートをリアルタイムで制御することにより表現性が向上するかを調査することを目的として実験を行った．また，パラメータ制御の遅延に関する客観評価及び操作性や遅延に関する主観評価の調査も行った．

4.1 比較条件

身体表現による制御の自然性を評価するため，ビブラートによる F_o の振幅が一定値である場合と 3.3 で挙げた 3 つの制御手法に対して，キーボードによる制御を加えた計 5 つの条件で実験を行った．なお，キーボードによる制御では 1 から 9 の 9 つのキーに離散的にパラメータ値が割り当てられ，このキーを押下中のみパラメータ制御を行うことができる．

a) 制御しない（ビブラートによる F_o 振幅が一定値）
b) キーボードによる制御
c) 手の上げ下げによる制御
d) 腕の伸ばし縮みによる制御
e) 眉毛の上げ下げによる制御

4.2 実験手順

実験は，22-26 歳の健常男性 10 名に対して行った．被験者には，はじめに歌唱する音源を試聴させることで実験の際に付与するビブラートの位置や程度の例を提示した．次に，被験者は健常者であり EL の使用経験がないため，パラメータ制御を行わない条件において練習させることで習得する時間を設けた．被験者が EL の使用方法を十分に理解した後，各パラメータ制御手法によるビブラート制御を用いた歌唱実験を無作為な順序で行った．この際，制御手法の取得のための練習時間には制限を設けないものとし，ビブラートを付与する箇所の指定は行わなかった．なお，ビブラートを制御する際に耳からの情報のみで制御可能であるかを検討するため，各条件における実験は制御によるパラメータの変化度合いを目視する場合（図 6）と聴覚のみでパラメータ変化を把握する場合の計 2 回ずつである．

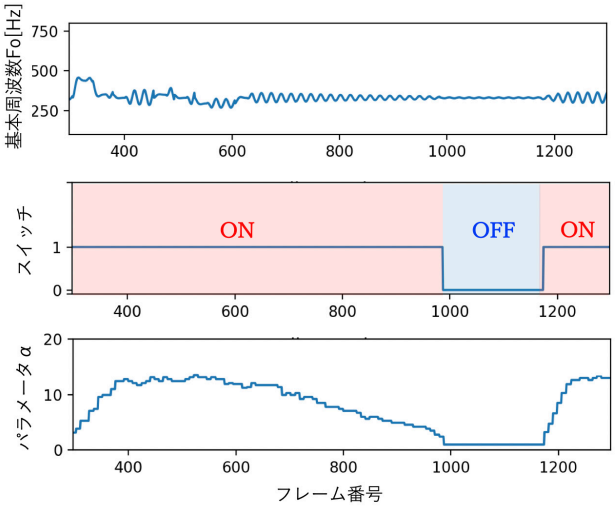

図 6 被験者が目視するビブラート制御のパラメータ情報．上段：基本周波数変動，中段：制御介入の有無，下段：ビブラート制御部に代入されるパラメータ

さらに実験終了後，付与されたビブラートの表現性に関する主観的な評価として以下の 1-2，提案するシステムにおける操作感に関する主観評価として 3-6 の計 6 項目における 5 段階評価（1：全くそう思わない－ 5：そう思う）への回答を求めた．

1. 付与されたビブラートは自然であると思う
2. 条件 a）と比較して表現性が向上した
3. 自然な身体表現でビブラートを付与できた（操作感なく付与することができた）
4. 意図した箇所にビブラートを付与することができた（ビブラートの反映に遅延を感じなかった）
5. ビブラートの強弱を意図通りに制御することができた
6. 使用の際パラメータ表示が必要だと感じた

5. 結果

5.1 ビブラート制御を用いた F_o 変動

キーボード制御による F_o 変動（図 7）では，パラメータ α を見ると被験者の意図した箇所にビブラートを付与できる一方で，変化が急峻であり，実際の歌唱における F_o 変動とは異なる結果が得られた．次に，腕の伸ばし縮みを用いた制御による F_o 変動（図 8）ではキーボード同様に意図通りの制御が可能であり，さらに滑らかにパラメータ α が変動している．最後に眉毛の上げ下げを用いた制御（図 9）による F_o 変動は制御が困難であり，ビブラートを付与したいと考えた箇所のパラメータ値が大きくなるものの，意図しない箇所においても付与される傾向が見られた．

5.2 主観評価実験

付与されたビブラートの表現性に関する評価結果は図 10 に示すように，付与された全ての制御手法において，付与されたビブラートは自然であり，歌唱全体としての主観的

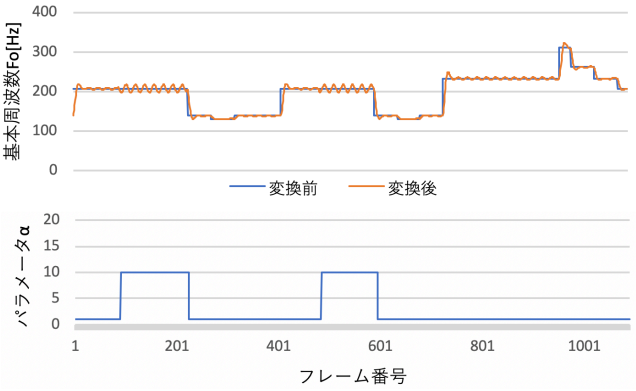

図 7 キーボード制御による F_o 変動

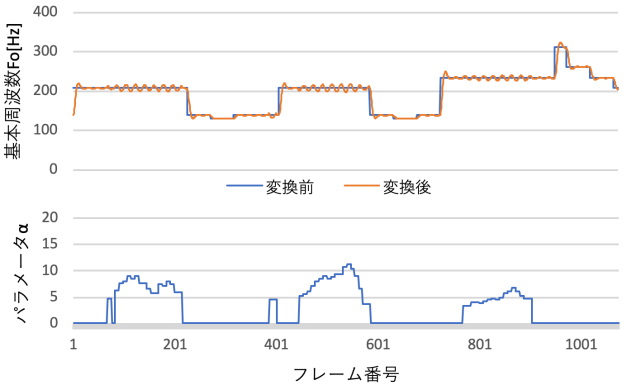

図 8 腕の伸ばし縮みを用いた制御による F_o 変動

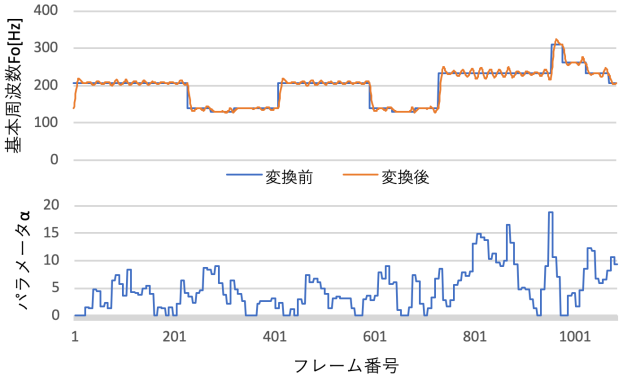

図 9 眉毛の上げ下げを用いた制御による F_o 変動

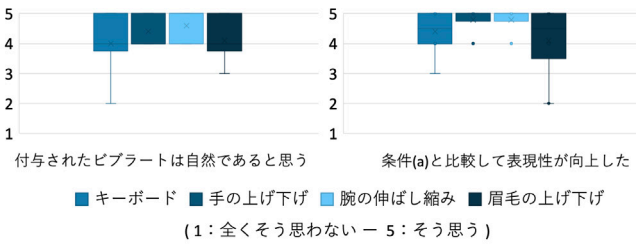

図 10 付与されたビブラートの表現性に対する主観評価

な表現性が向上することが分かった．また，提案するシステムにおける操作感に関する評価結果（図 11）を見ると，ビブラートを付与する際の身体表現についてはキーボードでの操作は不自然であり，付与する箇所や強弱のような操作性については 5.1 同様に，表情による操作が困難である

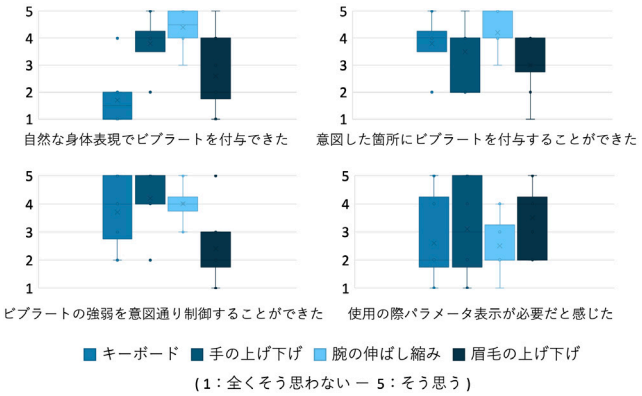

図 11 提案システムの操作性に関する主観評価

という傾向が見られる．一方で，骨格推定を用いた腕による制御では操作が容易であり，さらに意図通りに制御することができるということが分かった．

6. 考察

6.1 提案システムの操作性と遅延

まず，キーボードによる制御では，意図した箇所，強弱でビブラートを付与することができる一方で，操作感のある動作であるという回答が得られた．キーボードによる操作は簡潔で習得することが容易である一方，歌唱の際には現れるはずの動作ではないため妥当な結果だと言える．次に，骨格推定技術を用いた 2 つの制御においては自然な身体表現で意図通りに制御することができるが，手の上げ下げによる制御よりも腕の伸ばし縮みによる制御の方が操作による遅延を感じづらいという違いが生じた．この原因としては手の上げ下げによる制御は腕の伸ばし縮みによる制御よりも慣れない身体動作であったことが考えられる．最後に眉毛の上げ下げによる制御では，身体表現の自然性，操作性の双方において骨格推定技術を用いた制御よりも劣る結果となった．これは眉毛による制御では繊細なパラメータ調整が困難であり，さらに表情というものは，意思に関わらず歌唱中に自然に変化してしまうためだと考えられる．

次に，遅延についての評価をする．システムの遅延は身体表現の処理時間が約 75 ms，座標情報を音高制御のパラメータへと変換する処理時間が約 75 ms，電気音声変換の処理時間が 10 ms である．そのため，身体表現が歌唱表現に反映されるまでの定量的な遅延は約 160 ms である．一方で，主観評価の結果を見ると全ての条件においてほとんどの被験者が遅延を感じず，腕の伸び縮みによる制御が最も意図通りの箇所に付与できたという結果が得られた．これは，この制御手法には「腕を伸ばし始める」予備動作があり，この動作がパラメータ制御が ON になる閾値であるためだと考えられる．

最後に，パラメータ表示の必要性については被験者ごと

に回答が大きく異なり，有意な違いは見られなかった．一方で，自然性や操作性において肯定的な結果が得られた腕の伸ばし縮みについては他の 3 条件と比較して表示の必要がないという傾向が強いことから，用いる制御手法が簡潔であり，一定期間の訓練時間がある場合にはパラメータによる表示が不要になると考えられる．

6.2 リアルタイム F_o 制御による表現性の向上

図 7 に示す通り，眉毛による制御では意図通りの箇所にビブラートを付与することができなかった一方で，主観評価（図 10）ではビブラートの自然性及び，条件 a）と比較した表現性の双方において全ての条件でリアルタイムでの F_o 制御が有意に働くことが分かった．これは繊細な制御は不可能であっても歌唱者が歌声に介入すること自体が主観的な歌唱の表現性を向上させる上で大きな役割を担うためであると考えられる．一方で，本システムにおけるビブラートの振幅は 3.2 で示すように，定常振動モデルが選択された時のみにしか反映されないため，眉毛による操作における意思に反したパラメータ変動が歌声の出力に現れなかったこともこの結果の原因として挙げられる．

7. 今後の展望

本研究ではリアルタイムでユーザの意思を反映した歌声の制御を行うこと，そして，自然な身体表現を用いることで歌唱の表現性を向上させることを目的としてビブラートの F_o 振幅を 4 種類の制御方法で操作するという実験を行った．結果，眉毛のような顔の一部を用いた操作は自然であるものの程度の制御が難しく意図しない動作が生じる一方で，腕による制御は操作性も自然性も有意に良い傾向が見られた．今後の展望として筋電位などの映像以外を用いた制御手法，そして，本実験では行うことのできなかったグロウルやウィスパーのような声質をリアルタイムで制御する手法の検討をする予定である．

8. デモンストレーションの説明

デモンストレーションでは会場に実験で使用したものと同様の機器を設置する．来場者は EL を用いて歌唱し，その際，腕を伸ばしたり眉毛の上げ下げを行ったりすることで自然な身体表現によるビブラート表現の制御を体験する．

謝辞 本研究の一部は JST，CREST，JPMJCR19A3 の支援を受けたものである．

参考文献

[1] Morikawa, K. and Toda, T.: Electrolaryngeal speech modification towards singing aid system for laryngectomees, *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, pp. 610–613 (2017).

[2] Li, L., Toda, T., Morikawa, K., Kobayashi, K. and Makino, S.: Improving Singing Aid System for Laryngectomees With Statistical Voice Conversion and VAE-SPACE., *ISMIR*, pp. 784–790 (2019).

[3] Tanaka, K., Toda, T., Neubig, G., Sakti, S. and Nakamura, S.: A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion., *INTERSPEECH*, pp. 3067–3071 (2013).

[4] 姫野剛至, 田中裕人, 水町光徳, 松井謙二, 中藤良久：LPC 残差駆動歌唱システムにおける音韻性除去による音質改善の検討 (2016).

[5] Malathi, P., Sureshw, G. and Moorthi, M.: Enhancement of electrolaryngeal speech using Frequency Auditory Masking and GMM based voice conversion, *2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, IEEE, pp. 1–4 (2018).

[6] 中村圭吾，戸田智基，　猿渡洋，鹿野清宏：肉伝導人工音声の変換に基づく喉頭全摘出者のための音声コミュニケーション支援システム，Vol. 90, No. 3, The Institute of Electronics, Information and Communication Engineers, pp. 780–787 (2007).

[7] 田中宏, 戸田智基, グラム・ニュービック, サクリアニ・サクティ，　中村哲：リアルタイム音源予測に基づく電気式人工喉頭制御の実装，聴覚研究会資料= Proceedings of the auditory research meeting，Vol. 45, No. 4, 日本音響学会，pp. 251–256 (2015).

[8] Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y. and Sheikh, Y.: Supervision-by-Registration: An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 360–368 (2018).

[9] Kopaczka, M., Schock, J. and Merhof, D.: Super-realtime facial landmark detection and shape fitting by deep regression of shape model parameters (2019).

[10] 佐々木康輔，渡邉健斗，　橋本学，長田典子：顔キーポイントの移動方向コードに基づく個人差の影響を受けにくい表情認識，Vol. 138, No. 5, 一般社団法人 電気学会，pp. 611–618 (2018).

[11] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299 (2017).

[12] Schneider, P., Memmesheimer, R., Kramer, I. and Paulus, D.: Gesture recognition in RGB videos using human body keypoints and dynamic time warping, *Robot World Cup*, Springer, pp. 281–293 (2019).

[13] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z. et al.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015).