Vol.2021-MUS-131 No.27
Vol.2021-SLP-137 No.27
2021/6/18

固有声変換法を用いた重唱における調和度制御に関する検討

菊地 晏南[1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]

概要：本稿では，2 人の歌唱者による重唱について，その調和の度合い（調和度）と各歌唱者の声質の関係についての基礎検討を行った．歌声合成の研究が盛んに行われている一方，合唱音声合成の際は各歌唱者の独唱音声を単に重ね合わせる場合が多く，各歌唱者の声質の調和を考慮して合唱音声を生成することは少ない．しかし，実際の合唱では各歌唱者の声質は独唱時と異なる場合が多い．したがって本稿では，重唱の調和度と声質の関係に着目し，声質変換を用いて重唱音声を生成した．固有声変換法を用いて重唱を構成する各歌唱者の声質を変化させ，クラウドソーシングサービスを用いた主観聴取実験で重唱の調和度の変化を調べた．また，声楽の専門家 2 人に対して，重唱音声と重唱を構成する歌唱者の独唱音声の主観聴取実験を行った．その結果，重唱に適した声質の存在が示唆された．

1. はじめに

日本の音楽教育では合唱が扱われることが多く，合唱は我々にとって学生の頃から馴染み深い演奏形態の一つである．さらに，合唱を趣味とする社会人から構成される合唱団もあり，合唱はプロ，アマチュア関係なく楽しめ，多くの人と関わりを持つコミュニケーションの場ともなっている．

合唱団はよりよく調和する合唱を目指して，声の大きさを調節したり発声のタイミングを合わせるよう練習に励む．このとき，各歌唱者は練習を通して他の歌唱者の声を聞き，自らの声質を調整することで互いの声がよく混ざり合うように発声する．結果として各歌唱者の合唱時の声質が独唱時と異なる場合が多く，このことは合唱を経験したことがあれば実感できるだろう．さらに，プロ歌唱者とアマチュア歌唱者を対象として合唱時と独唱時の音声を比較した研究では，合唱時と独唱時でスペクトルに異なる特徴が見られた [1]．しかし，合唱時に歌唱者が具体的にどのような声を目指して歌唱すれば，合唱団全体としてより調和した合唱音声を生み出すことができるかを正確に記述することは難しい．したがって，各歌唱者の独唱音声から，それらを単に重ね合わせた音声よりも調和した合唱音声を生成したり，そのための各歌唱者の目指すべき歌唱音声を生成したりできれば，合唱練習時の指針となると考えられる．

[1] 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
[a)] kikuchi.1218@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp

各歌唱者の独唱音声から合唱音声を生成する合唱制作支援インタフェースとしては，Unisoner が提案されている [2]．Unisoner では， 同じ楽曲を複数の歌唱者がそれぞれ歌った Web 上の音声を用いて，各歌唱音声の歌声以外の伴奏音声を抑制し，各歌唱音声の波形を切り取って重ね合わせることで合唱音声を生成する．ただし，各歌唱音声の音量や左右 2 チャネル出力における左右の音量は調節できるが，各歌唱音声を単に重ね合わせるだけで声質の調整は行わない．

そこで本稿では，2 人による重唱[*1]において，声質変換技術を用いて各歌唱者の声質を変化させることで，重唱の調和の度合い（調和度）を制御する方法について基礎的な検討を行う．本稿では，声質変換技術として固有声を用いた一対多声質変換 [3] を用いた．これにより，生成される歌唱音声における発声タイミングや基本周波数のばらつきを抑制し，声質と調和度の関係を調べることができる．固有声を用いた一対多声質変換により主旋律と副旋律の独唱音声を生成し，それらを重ね合わせることで重唱音声を生成した．生成した重唱音声に対してクラウドソーシングサービスを用いた主観聴取実験と，重唱音声とそれを構成する主旋律と副旋律それぞれの独唱音声に対して声楽の専門家による主観聴取実験を行った．その結果，重唱に適した声質をもつ歌唱者が存在する可能性が確認された．

2. 複数歌唱者の調和制御

本節では，ソースフィルタ分解に基づいて 2 人による重唱音声の調和制御の検討を行った研究 [4] について述べる．

[*1] 複数のパートをそれぞれ複数の歌唱者が歌う場合を合唱，複数のパートをそれぞれ 1 人の歌唱者が歌う場合を重唱という．

なお，本稿では 1 節で述べたように調和の度合いを「調和度」と呼ぶが，[4] では調和度という表現は用いられていないため，本節および [4] に関する記述においては調和という表現を用いる．

複数歌唱者の調和制御に関する研究では，2 人による重唱において，ペアになる歌唱者の違いが重唱の調和に与える影響についての検討がなされた [4]．まず，重唱を構成する歌唱者 2 人について，各独唱音声のメルケプストラム系列を動的時間伸縮 (Dynamic time warping; DTW) によって対応づける．次に，それらのメルケプストラム歪み (Mel-cepstral distortion; MCD) を求め，これをペア内の MCD とし，ペア内の MCD が異なる 11 ペアの重唱音声について自然性を評価した．ここで，重唱の自然性は，主観聴取実験によってより自然だと感じた音声を選択する総当たりの AB テストを行い，その結果をサーストンの一対比較法によって間隔尺度とした値を用いて評価した．また，MCD は 2 つのメルケプストラムの距離尺度であり，以下の式で表される．

$$\mathrm{MCD}\ [\mathrm{dB}] = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{D_{mc}}\left(mc_d^{(1)} - mc_d^{(2)}\right)^2} \qquad (1)$$

ただし，$mc_d^{(1)}$，$mc_d^{(2)}$ は各メルケプストラムの d 次の係数，D_{mc} はメルケプストラムの最大次数である．

聴取実験の結果，ペア内の MCD と重唱の自然性に相関が見られ，特にペア内の MCD が 7.5 [dB] を超えると重唱の自然性が低くなる傾向が見られた．

3. 固有声

話者認識における話者適応手法の一つとして，固有声を用いた手法が提案された [5]．固有声は顔認識における固有顔に着想を得たものである．固有声を用いた話者適応手法では，話者適応後のモデルのパラメータを少量の基底ベクトルの線形結合で表現する．したがって，話者適応時は基底ベクトルの係数にあたる重みベクトルのみを推定するため，推定するパラメータ数が大幅に削減でき，適応に用いる発話データの数を抑えることができる．

固有声を用いた話者適応手法における基底ベクトルは固有声と呼ばれ，以下の手順によって求める．まず，事前学習用話者 S 人分の話者依存モデルと，1 つの話者非依存モデルを構築する．次に，S 人分の話者依存モデルから D_{SV} 次元のスーパーベクトルをそれぞれ抽出する．ここで，スーパーベクトルは推定されるモデルパラメータを並べたものである．S 個の D_{SV} 次元スーパーベクトルに次元削減法を適用することで，S 個の D_{SV} 次元基底ベクトルを得る．固有声として K 個の基底ベクトルを求める際は，これらの基底ベクトルのうち，事前学習用話者間変動への寄与の大きいものから順に K 個を選択する．こうして得られた K 個の基底ベクトルを固有声と呼ぶ．

話者適応の際は，適応話者の発話データから固有声空間上の座標に相当する K 次元の重みベクトルを推定し，これを用いてモデルパラメータを推定することで，適応話者の話者依存モデルを得る．

4. 固有声を用いた一対多声質変換

本稿では，固有声を用いた一対多声質変換 [3] によって重唱音声を生成した．一対多声質変換は，ある話者の音声を任意の話者の音声に変換する技術である．

固有声を用いた一対多声質変換における処理は，固有声混合ガウスモデル (Eigenvoice Gaussian Mixture Model; EV-GMM) を事前学習する処理，入力話者と出力話者のデータで EV-GMM を適応させる適応処理，適応させた EV-GMM を用いて入力話者の音声を変換する処理から成る．

4.1 EV-GMM の事前学習処理

まず，入力話者と多数の事前学習用出力話者のパラレルデータで 話者非依存 GMM $\boldsymbol{\lambda}^{(0)}$ を事前学習する．次に，入力話者と s 番目の事前学習用出力話者のパラレルデータを用いて話者非依存 GMM $\boldsymbol{\lambda}^{(0)}$ の平均ベクトルのみを更新することで，s 番目の事前出力話者の話者依存 GMM $\boldsymbol{\lambda}^{(s)}$ を学習する．ここで，事前学習用出力話者数を S として，$s = 1, 2, ..., S$ である．学習した話者依存 GMM $\boldsymbol{\lambda}^{(s)}$ の各正規分布の平均ベクトルを接続することで，スーパーベクトル $\boldsymbol{SV}^{(s)} = [\boldsymbol{\mu}_1^{(Y,s)\top}, \boldsymbol{\mu}_2^{(Y,s)\top}, ..., \boldsymbol{\mu}_M^{(Y,s)\top}]^\top$ を得る．全事前学習用出力話者に対するスーパーベクトルに対して主成分分析 (Principal component analysis; PCA) を行うことで，バイアスベクトル $\boldsymbol{b}^{(0)} = [\boldsymbol{b}_1^{(0)\top}, \boldsymbol{b}_2^{(0)\top}, ..., \boldsymbol{b}_M^{(0)\top}]^\top$ と固有ベクトルから構成される $\boldsymbol{B} = [\boldsymbol{B}_1^\top, \boldsymbol{B}_2^\top, ..., \boldsymbol{B}_M^\top]^\top$ を決定する．

$$\boldsymbol{SV}^{(s)} \approx \boldsymbol{B}\boldsymbol{w}^{(s)} + \boldsymbol{b}^{(0)} \qquad (2)$$

$$\boldsymbol{b}_m^{(0)} = \frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\mu}_m^{(Y,s)} \qquad (3)$$

ここで，$\boldsymbol{w}^{(s)}$ は s 番目の事前学習用出力話者に対する K 次元の重みベクトルである．EV-GMM $\boldsymbol{\lambda}^{(EV)}$ は，$\boldsymbol{b}^{(0)}$，$\boldsymbol{B}$ および $\boldsymbol{\lambda}^{(0)}$ で構成される．

4.2 適応処理

入力話者の静的・動的特徴量系列，出力話者の静的・動的特徴量系列に対して DTW を用いて対応をとり，それぞれを

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_T]^\top \qquad (4)$$

$$\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_T]^\top \qquad (5)$$

とする．ここで，$^\top$ は転置を表し，T はフレーム数である．

出力話者に対応する重みベクトル $\hat{\boldsymbol{w}}$ は，EV-GMM

の教師なし適応に MLED (Maximum likelihood eigen-decomposition) [5] を用いることによって次のように推定する．

$$\hat{\boldsymbol{w}} = \arg\max \int P\left(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\lambda}^{(EV)}\right) d\boldsymbol{X} \tag{6}$$

$$= \arg\max \int P\left(\boldsymbol{Y}|\boldsymbol{\lambda}^{(EV)}\right) P\left(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(EV)}\right) d\boldsymbol{X} \tag{7}$$

$$= \arg\max P\left(\boldsymbol{Y}|\boldsymbol{\lambda}^{(EV)}\right) \tag{8}$$

確率密度関数のモデルとして GMM を用いているので，次の補助関数 $Q(\boldsymbol{w}, \hat{\boldsymbol{w}})$ を繰り返し最大化することで $\boldsymbol{w}$ を求める．

$$Q(\boldsymbol{w}, \hat{\boldsymbol{w}}) = \sum_{m=1}^{M} P\left(m|\boldsymbol{Y}, \boldsymbol{\lambda}^{(EV)}\right) \log P\left(\boldsymbol{Y}, m|\hat{\boldsymbol{\lambda}}^{(EV)}\right) \tag{9}$$

よって，$\hat{\boldsymbol{w}}$ は

$$\hat{\boldsymbol{w}} = \boldsymbol{D}^{-1}\boldsymbol{N} \tag{10}$$

と表される．ここで，

$$\boldsymbol{N} = \sum_{m=1}^{M} \boldsymbol{B}_m^{\top} {\boldsymbol{\Sigma}_m^{(YY)}}^{-1} \bar{\boldsymbol{Y}}_m \tag{11}$$

$$\boldsymbol{D} = \sum_{m=1}^{M} \bar{\gamma}_m \boldsymbol{B}_m^{\top} {\boldsymbol{\Sigma}_m^{(YY)}}^{-1} \boldsymbol{B}_m \tag{12}$$

$$\bar{\gamma}_m = \sum_{t=1}^{T} P\left(m|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(EV)}\right) \tag{13}$$

$$\bar{\boldsymbol{Y}}_m = \sum_{t=1}^{T} P\left(m|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(EV)}\right) \left(\boldsymbol{Y}_t - \boldsymbol{b}_m^{(0)}\right) \tag{14}$$

である．また，$\boldsymbol{\Sigma}_m^{(YY)}$ は m 番目の正規分布における出力話者の分散共分散行列である．

4.3 変換処理

変換の際は，学習された EV-GMM $\boldsymbol{\lambda}^{(EV)}$ と推定した出力話者に対応する重みベクトル $\hat{\boldsymbol{w}}$ を用いて従来の GMM を用いた声質変換と同様に変換を行う．ただし，$\boldsymbol{X}_t$ が与えられた時の m 番目の条件付き確率密度分布の平均ベクトル $\boldsymbol{E}_{m,t}^{(Y)}$ は

$$\boldsymbol{E}_{m,t}^{(Y)} = \hat{\boldsymbol{\mu}}_m + \boldsymbol{\Sigma}_m^{(YX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right) \tag{15}$$

と表される．ここで，

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{B}_m \hat{\boldsymbol{w}} + \boldsymbol{b}_m^{(0)} \tag{16}$$

である．また，$\boldsymbol{\Sigma}_m^{(XX)}$，$\boldsymbol{\Sigma}_m^{(YX)}$ は m 番目の正規分布における入力話者の分散共分散行列，出力話者と入力話者の相互共分散行列であり，$\boldsymbol{\mu}_m^{(X)}$ は m 番目の正規分布における入力話者の平均ベクトルである．

5. 固有声を用いた重唱の調和度制御

本稿では，固有声を用いた一対多声質変換によって声質と重唱の調和度の関係を調査する．複数歌唱者の調和制御の研究では，メルケプストラム空間上で重唱を構成する各歌唱者を近づけることによる調和制御を検討した [4]．しかし，メルケプストラム空間はあらゆる音響現象を記述するため，音声を対象とするためのより適切な空間を用いた場合についての検討も考えられる．そこで本稿では，固有声空間上で歌唱者の声質を変換する．固有声空間は事前学習時に PCA によって事前学習用出力話者の GMM のスーパーベクトルを次元削減するため，話者の違いを表現するのに適している．

また，複数歌唱者の調和制御の研究では歌唱者間の距離を MCD として歌唱者間距離と重唱の自然性の関係を調査した [4]．しかし，様々なペアを用いて重唱を生成したため，声質の違い以外にも，各音を発するタイミングのずれなどの各歌唱者の歌い方の違いが聴取実験の結果に影響している可能性がある．そこで，一対多声質変換を用いることで，歌い方の影響を除去し，声質の比較が可能となる．

6. 実験条件

本稿では，歌唱データとして JVS-MuSiC [6] を用いた．JVS-MuSiC には男性 49 人，女性 51 人の歌唱音声が収録されている．各歌唱者ごとに 2 曲が収録されており，全歌唱者共通の童謡「かたつむり」と，各歌唱者ごとに異なる童謡 1 曲である．共通曲に関しては，各歌唱者がピッチとテンポを自由に決めて歌った音声，各歌唱者ごとにピッチとテンポを統一した音声，グループごとにピッチとテンポを統一した音声の 3 種類が収録されている．グループとは，男女をそれぞれ 3 グループに分けた計 6 グループのことであり，各グループ内でピッチとテンポが統一されている．本実験ではこのうち女性歌唱者 51 人の，グループごとにピッチとテンポが統一された音声を使用した．JVS-MuSiC には「かたつむり」が 2 番まで収録されているが，聴取実験では 1 番のみを使用し，音声の長さは約 15 秒である．

重唱音声を生成するにあたって使用した楽譜を**図 1** に示す．この楽譜において，音高が高いパートを主旋律とし，音高が低いパートを副旋律とする．主旋律は JVS-MuSiC に収録された音声から得られた F0 系列全体を楽譜に従ってスケーリングすることで得られるが，副旋律は曲全体での一律のスケーリングでは得られないため，[4] で用いた副旋律生成手順に則り，以下のように合成した．まず，Julius [7][8] を用いて強制アライメントを取った後，[4] の著者が手動で修正して時系列音素ラベルを得た．この音素ラベルと図 1 に示した楽譜を用いて F0 系列を生成し，さらに，人間の歌唱として自然な音声にするために F0 系列に人間の歌唱に特徴的な動的変動成分 [9] を付与することで副旋

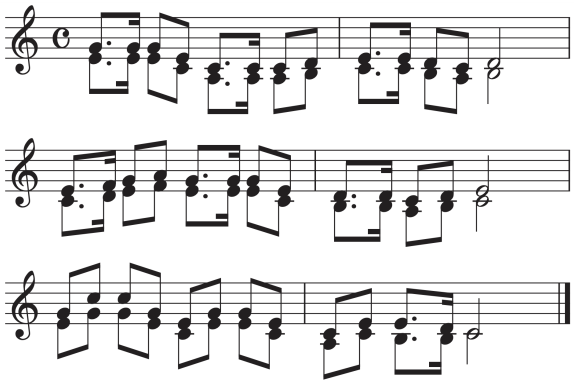

図 1 実験で用いた童謡「かたつむり」の楽譜 [4]

表 1 実験で用いられたペア．歌唱者 ID は各歌唱者の属するクラスタ番号と，‘‘a’’ または ‘‘b’’ からなる．‘‘a’’ は代表歌唱者を，‘‘b’’ は副代表歌唱者を表す．

	Main Part		Harmony Part	
Pair ID	Singer ID	Singer	Singer ID	Singer
A	1b	jvs055	1a	jvs069
B	2a	jvs016	2b	jvs039
C	3b	jvs007	3a	jvs017
D	4b	jvs015	4a	jvs082
E	2a	jvs016	1a	jvs069
F	3a	jvs017	1a	jvs069
G	4a	jvs082	1a	jvs069
H	3a	jvs017	2a	jvs016
I	2a	jvs016	4a	jvs082
J	3a	jvs017	4a	jvs082

律を合成した．

実験に用いる重唱音声は，固有声を用いた一対多声質変換によって生成した主旋律と副旋律の重ね合わせである．すなわち，変換元歌唱者の JVS-MuSiC に収録されている音声の F0 をスケーリングして得た主旋律と，上記の手順で得た副旋律の音声をそれぞれ変換元音声として，定められたペアの主旋律と副旋律の音声を生成し，それらの重ね合わせを重唱音声とした．変換元となる歌唱者としては，JVS-MuSiC の女性歌唱者 51 人の中からランダムに選んだ jvs051 を用いた．声質変換は最尤規準で推定した入出力差分特徴量を用いて差分スペクトル法 [10] により行った．

本稿では，重唱ペアとして**表 1** の 10 ペアを用いた．ここで，表 1 の 10 ペアは以下のように選定した．まず，女性歌唱者 51 人から変換元歌唱者である jvs051 を除いた 50 人の固有声の重みベクトルを推定し，ward 法によって重みベクトルのクラスタリングを行った．クラスタリング結果のデンドログラムを**図 2** に示す．なお，クラスタ数は 4 とし，各クラスタ名はクラスタ内で最後に結合した時の距離が小さい順にクラスタ 1 からクラスタ 4 とした．次に，これら 4 クラスタの重心を求め，各クラスタに属する重みベクトルの中からクラスタの重心ベクトルとのユークリッド距離が最も小さい重みベクトルを選出し，その重みベクトル

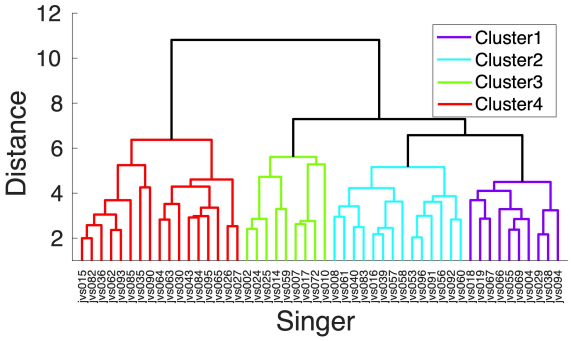

図 2 女性歌唱者 50 人の固有声の重みベクトルを ward 法でクラスタリングした際のデンドログラム

ルの歌唱者をそのクラスタの代表歌唱者とした．各クラスタの代表歌唱者と同じクラスタに属し，かつ代表歌唱者の重みベクトルとのユークリッド距離が最も小さい重みベクトルを持つ歌唱者をそのクラスタの副代表歌唱者とした．こうして各クラスタから選出された代表歌唱者，副代表歌唱者を用いて，各クラスタでクラスタ内ペアを 1 つ，各クラスタの代表歌唱者の総当たりでクラスタ間ペアを 6 つ作成した．また，重唱ペアのパート分けに関しては，50 人の女性歌唱者の固有声の重みベクトルの重心を求め，重心からのユークリッド距離がより大きい重みベクトルを持つ歌唱者を主旋律とした．

音声の分析合成には WORLD [11]（D4C edition [12]）を用いた．音声のサンプリング周波数は 24 kHz，FFT の窓長は 1024 サンプル（約 43 ms），シフト長は 5 ms である．メルケプストラムの次数は 25 次であり，GMM の学習には静的・動的特徴量を用いた．

本稿では，生成した音声に対し，Web 上のクラウドソーシングサービスを用いた主観聴取実験と，声楽の専門家による主観聴取実験を行った．

7. 実験 1：クラウドソーシングサービスを用いた評価

7.1 実験条件

表 1 における 10 ペアの重唱に対し，Web 上のクラウドソーシングサービスを用いて主観聴取実験を行った．各重唱音声に対し，合唱として，より自然だと感じられる音声を選択する AB テストを総当たりで行い，その結果からサーストンの一対比較法によって間隔尺度を算出した．こうして得られた間隔尺度を本稿では重唱の調和度の間隔尺度とする．被験者数は一対の比較に対して 25 人であり，順序効果を考慮して各被験者は一対の比較について 2 回の評価を行った．

Vol.2021-MUS-131 No.27
Vol.2021-SLP-137 No.27
2021/6/18

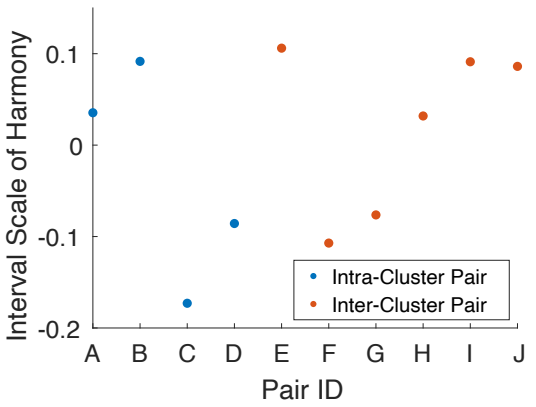

図 3 歌唱者のクラスタリング結果に基づいて決定したペアの重唱の調和度の間隔尺度

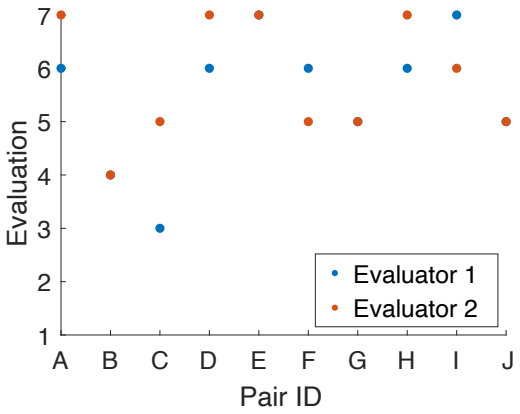

図 4 声楽の専門家による重唱音声の 7 段階評価

表 2 実験 2 で用いられた 7 段階の評価基準

	重唱 [独唱] の評価基準
1	全く調和していない [適していない]
2	調和していない [適していない]
3	どちらかというと調和していない [適していない]
4	どちらとも言えない
5	どちらかというと調和している [適している]
6	調和している [適している]
7	十分調和している [適している]

7.2 実験結果

歌唱者のクラスタリング結果に基づいて決定したペアの，重唱の調和度の間隔尺度を**図 3** に示す．図 3 より，ペア内にクラスタ 2 の歌唱者がいるペア B，E，H，I はその他のペアに比べ，調和度が高い傾向がある．このことから，クラスタ 2 の代表歌唱者である歌唱者 ID が 2a の歌唱者の声質は他の歌唱者に比べて重唱に適している可能性があると考えられる．また，クラスタ 3，4 に着目すると，クラスタ 3，4 間ペア J は調和度が高く，それぞれのクラスタ内ペア C，D とクラスタ 1 とのペア F，G は比較的調和度が低く，クラスタ 2 とのペア H，I は調和度が高い．このことから，クラスタ 3，4 の代表歌唱者は重唱の調和度に関して同様の傾向を有する可能性があると考えられる．ただし，クラスタ全体の傾向を論じるには本実験よりもペア数を増やす必要がある．

8. 実験 2：声楽の専門家による評価

本実験では，7 節で用いた重唱音声に対して声楽の専門家 2 人による評価実験を行った．また，これらの重唱音声を合成する前段階で合成した主旋律と副旋律それぞれの独唱音声についても同じ専門家による評価実験を行った．重唱音声に対してはどの程度調和しているかを，独唱に対しては声質がどの程度重唱に適しているかを，**表 2** に示す 7 段階で評価した．

また，それぞれの音声に対してその評価理由を自由に記述してもらった．評価の際は重唱音声全てまたは独唱音声全てを提示し，各音声を再生する回数は無制限とした．

重唱音声に対する評価の結果を**図 4** に示す．クラスタ 2 の歌唱者が含まれるペア B，E，H，I のうち，ペア E，H，I は他の重唱と比べて調和度が高かった．

次に，独唱音声に対する評価の結果を**図 5** に示す．図 5 より，クラスタ 2 の代表歌唱者である歌唱者 ID が 2a の歌唱者は，2 人の評価者いずれからも「7：十分適している」と評価された．また，歌唱者 ID が 2a の歌唱者の主旋律の独唱は，評価者 1 からは「副旋律を選ばずハーモニーが聞こえそうだ」，評価者 2 からは「重唱にした時に合うものの振り幅が大きいと感じる」と評価された．したがって，歌唱者 ID が 2a の歌唱者は重唱に適した声質である可能性が示唆された．ここで，歌唱者 ID が 2a の歌唱者が主旋律を歌唱するペアは，ペア ID が B，E，I の重唱である．したがって，図 4 においてペア ID が E，I の重唱の評価が比較的高かった理由として，歌唱者 ID が 2a の歌唱者の寄与が考えられる．なお，ペア ID が B の重唱は，評価者 1 からは「副旋律がやや強い」，評価者 2 からは「副旋律が鋭く感じられる」と評価された．よって，ペア ID が B の重唱が E，I の重唱に比べて評価が低かった理由は，主旋律歌唱者である 2a の歌唱者ではなく，副旋律歌唱者である可能性が考えられる．

9. おわりに

本稿では，話者適応の分野で導入された固有声を声質変換に応用した固有声変換法 [3] を用いて，2 人による重唱音声を生成し，重唱の調和と重唱を構成する歌唱者の声質の関係について調査した．歌唱者を凝集型クラスタリングの一つである ward 法を用いてクラスタリングし，各クラスタを代表する歌唱者を用いて重唱音声を生成し，重唱の調和度とクラスタとの関係を調べた．生成した重唱音声に対して，クラウドソーシングサービスを用いた主観聴取実験を行った結果，重唱に適した声質を持つ歌唱者の存在が示唆された．また，これらの重唱音声と重唱を構成する主

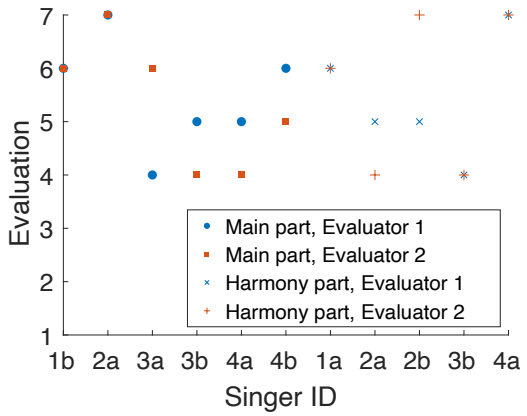

図 5 声楽の専門家による独唱音声の 7 段階評価

旋律と副旋律それぞれの独唱音声に対して，声楽の専門家による主観聴取実験を行った．その結果，クラウドソーシングサービスを用いた主観聴取実験において重唱に適した声質を持つ可能性が示唆された歌唱者は，幅広い副旋律歌唱者との重唱で調和しそうであると評価された．

本稿でのモデルの学習や聴取実験には，JVS-MuSiC に収録されている全歌唱者が共通で歌った童謡「かたつむり」のみを用いた．しかし，JVS-MuSiC には本稿で用いた童謡「かたつむり」以外に，各歌唱者ごとに異なる童謡の歌唱音声も収録されている．したがって，これらを用いてより多くの曲で聴取実験を行うことが可能である．また，本稿では互いに類似した歌唱者同士が同じクラスタに属することを期待して，固有声の重みベクトルのユークリッド距離で歌唱者を 4 つのクラスタにクラスタリングした．しかし，クラスタ内で声質が類似することと，そのために最適なクラスタ数は自明ではない．本稿でクラスタ数を 4 としたのは聴取実験の規模を考慮したものであったが，異なるクラスタ数で声質の類似度や重唱の調和度に関する傾向を調査する必要がある．さらに，歌唱者表現として固有声の他に i-vector [13] などの他の話者埋め込みを用いることも考えられる．

参考文献

[1] Rossing, T., Sundberg, J. and Ternström, S.: Acoustic comparison of voice use in solo and choir singing, *The Journal of the Acoustical Society of America*, Vol. 79, No. 6, pp. 1975–1981 (1986).
[2] 都築圭太，中野倫靖，後藤真孝，山田武志，牧野昭二：Unisoner：様々な歌手が同一楽曲を歌った Web 上の多様な歌声を活用する合唱制作支援インタフェース，情報処理学会論文誌, Vol. 56, No. 12, pp. 2370–2383 (2015).
[3] Toda, T., Ohtani, Y. and Shikano, K.: Eigenvoice Conversion Based on Gaussian Mixture Model, *Proc. INTERSPEECH-2006*, pp. 2446–2449 (2006).
[4] 山内孔貴，須田仁志，齋藤大輔，峯松信明：ソースフィルタ分解に基づく複数歌唱者の調和制御に関する検討，情報処理学会研究報告，Vol. 2020-SLP-132, No. 35, pp. 1–6 (2020).
[5] Kuhn, R., Junqua, J. C., Nguyen, P. and Niedzielski, N.: Rapid speaker adaptation in eigenvoice space, *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 6, pp. 695–707 (2000).
[6] Tamaru, H., Takamichi, S., Tanji, N. and Saruwatari, H.: JVS-MuSiC: Japanese multispeaker singing-voice corpus, *arXiv:2001.07044[cs.SD]* (2020).
[7] Lee, A., Kawahara, T. and Shikano, K.: Julius — An Open Source Real-Time Large Vocabulary Recognition Engine, *EUROSPEECH2001: the 7th European Conference on Speech Communication and Technology*, Vol. 3, No. 7, pp. 1691–1694 (2001).
[8] Lee, A.: Recent Development of Open-Source Speech Recognition Engine Julius, *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 131–137 (2009).
[9] Saitou, T., Unoki, M. and Akagi, M.: Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis, *Speech Communication*, Vol. 46, pp. 405–417 (2005).
[10] Kobayashi, K., Toda, T., Neubig, G., Sakti, S. and Nakamura, S.: Statistical singing voice conversion with direct waveform modification based on the spectrum differential, *Proc. INTERSPEECH-2014*, pp. 2514 – 2518 (2014).
[11] 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. E99-D, No. 7, pp. 1877–1884 (2016).
[12] Morise, M.: D4C, a band-aperiodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol. 84, pp. 57–65 (2016).
[13] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. and Ouellet, P.: Front-End Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798 (2011).