多対多固有声変換に基づく歌声声質変換及び歌声合成を用いた学習データ生成*

☆土井啓成，戸田智基 （奈良先端大），中野倫靖，後藤真孝 （産総研），
中村哲 （奈良先端大）

1 はじめに

歌声の声質には歌手の個人性が反映されており，他者の声質に自在に切り替えて歌うことは難しいが，それが可能になれば，歌唱の楽しさを増し，歌唱表現の可能性を広げることができる．そこで我々は，歌手の声質を他の歌手の声質へと変換することで，多様な声質での歌唱の実現を目指す．歌声の声質を変換する一手法として，源歌手と目標歌手の声質の対応関係を統計的に学習し，両歌手間の歌声の変換を行う歌声声質変換 [1] が提案されている．しかしながら，学習時に，源歌手と目標歌手が同一曲を歌唱した歌声（パラレルデータ）が必要となるため，そのような歌声が得られない歌手間での変換は不可能であった．本稿では，任意の歌手間での変換を容易に実現するために，多対多固有声変換 [2] に基づく歌声声質変換を提案する．さらに，多対多固有声変換における事前学習で必要な歌声収録の負担を減らすため，歌唱表現を模倣できる歌声合成システム VocaListener [3] を用いた学習データ生成法を提案する．

2 従来の歌声声質変換と歌声合成

源歌手の歌声を目標歌手の歌声へ統計的手法で変換する歌声声質変換 [1] は，学習処理と変換処理から成る．学習時には，源歌手と目標歌手が同一曲を歌唱した歌声で構成されるパラレルデータを用い，両歌手の音響特徴量の結合確率密度関数を混合正規分布モデル (Gaussian mixture model: GMM) でモデル化する．変換時には，新たに収録された源歌手の歌声を，GMM に基づき，最尤系列変換法 [4] を用いて目標歌手の歌声へと変換する．

歌声合成は，所望の声質を持つ歌声合成用コーパスを利用し，歌声を生成する技術であり，中でも，歌詞と楽譜情報から歌声を合成する VOCALOID2 [5] が有名である．VOCALOID2 では，音高や音量といった合成パラメータの手動操作により，自然な歌声を生成可能であるが，そのパラメータ操作には多大な労力が必要であった．この問題に対し，中野ら [3] によって，VOCALOID2 等の合成パラメータをユーザの手本歌声から自動推定し，手本歌声を模倣した表現力豊かな合成歌声を容易に生成できる歌声合成システム VocaListener が提案されている．

3 従来の話声に対する多対多固有声変換

学習データとして，参照話者と多数の事前収録目標話者による同一内容の発話を用いる．時間フレーム t における参照話者と s 人目の事前収録目標話者の静的・動的結合特徴量ベクトルをそれぞれ $\boldsymbol{X}_t$, $\boldsymbol{Y}_t^{(s)}$ とする．全ての事前収録目標話者に対する静的・動的結合特徴量ベクトル対を学習データとして用いて，固有声 GMM(Eigenvoice GMM) [6] を学習する．この時，$\boldsymbol{X}_t$ と $\boldsymbol{Y}_t^{(s)}$ の結合確率密度関数は次式にて表される．

$$P\left(\boldsymbol{X}_t,\boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EV)},\boldsymbol{w}^{(s)}\right) = \sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{Y}_t^{(s)}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(X)}\\ \boldsymbol{\mu}_m^{(Y,s)}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)}\\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (1)$$

ここで，全 M 個の分布中の m 番目の分布における s 人目の事前収録目標話者に対する平均ベクトル $\boldsymbol{\mu}_m^{(Y,s)}$ は，次式で与えられる．

$$\boldsymbol{\mu}_m^{(Y,s)}=\boldsymbol{B}_m^{(Y)}\boldsymbol{w}^{(s)}+\boldsymbol{b}_{m,0}^{(Y)} \quad (2)$$

行列 $\boldsymbol{B}_m^{(Y)} = [\boldsymbol{b}_{m,1}^{(Y)},\cdots,\boldsymbol{b}_{m,J}^{(Y)}]$ 及びベクトル $\boldsymbol{b}_{m,0}^{(Y)}$ は，m 番目の分布の基底ベクトルセット（ベクトル数は J）及びバイアスベクトルであり，$\boldsymbol{w}^{(s)} = \left[w^{(s)}(1),\cdots,w^{(s)}(J)\right]^\top$ は s 人目の事前収録目標話者に対する J 次元の重みベクトルである．一方，パラメータセット $\boldsymbol{\lambda}^{(EV)}$ は，個々の分布における混合重み α_m，参照話者の平均ベクトル $\boldsymbol{\mu}_m^{(X)}$，上記の $\boldsymbol{B}_m^{(Y)}$ と $\boldsymbol{b}_{m,0}^{(Y)}$，および，各共分散／相互共分散行列 $\boldsymbol{\Sigma}_m^{(XX)}$, $\boldsymbol{\Sigma}_m^{(XY)}$, $\boldsymbol{\Sigma}_m^{(YX)}$, $\boldsymbol{\Sigma}_m^{(YY)}$ から成り，全事前収録目標話者間で共有される．

適応処理では，任意の元話者 i に対する重みベクトル $\hat{\boldsymbol{w}}^{(i)}$ は次式により推定される．

$$\hat{\boldsymbol{w}}^{(i)} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\prod_{t=1}^{T}\int P\left(\boldsymbol{X}_t,\boldsymbol{Y}_t^{(i)}|\boldsymbol{\lambda}^{(EV)},\boldsymbol{w}\right)\mathrm{d}\boldsymbol{X}_t \quad (3)$$

ここで，$\boldsymbol{Y}_t^{(i)}$ は，時間フレーム t における元話者 i の音響特徴量の静的・動的結合特徴量ベクトルである．同様に，任意の目標話者 o の重みベクトル $\hat{\boldsymbol{w}}^{(o)}$ も独立に推定される．元話者 i と目標話者 o の音響特徴量の結合確率密度関数は，次式にて導出される．

$$P\left(\boldsymbol{Y}_t^{(i)},\boldsymbol{Y}_t^{(o)}|\boldsymbol{\lambda}^{(EV)},\hat{\boldsymbol{w}}^{(i)},\hat{\boldsymbol{w}}^{(o)}\right) = \sum_{m=1}^{M}\alpha_m\int P\left(\boldsymbol{Y}_t^{(i)}|\boldsymbol{X}_t,m,\boldsymbol{\lambda}^{(EV)},\hat{\boldsymbol{w}}^{(i)}\right) P\left(\boldsymbol{Y}_t^{(o)}|\boldsymbol{X}_t,m,\boldsymbol{\lambda}^{(EV)},\hat{\boldsymbol{w}}^{(o)}\right)P\left(\boldsymbol{X}_t|m,\boldsymbol{\lambda}^{(EV)}\right)\mathrm{d}\boldsymbol{X}_t \quad (4)$$

変換時には，最尤系列変換法に基づき，適応された固有声 GMM を用いて，元話者の発話を目標話者の発話へ変換する．

4 提案法

多対多固有声変換を歌声声質変換に適用する．提案法 [7] では，任意の源歌手と目標歌手による少量（例えば数秒程度）の歌声を使用して，固有声 GMM を適応することで，両歌手間の歌声声質変換を実現できる．適応に用いる曲は，両歌手で異なっていてもよい．このため，様々な目標歌手の声質への変換は極めて容易となり，システム使用時におけるユーザ（源歌手）の事前準備の手間は大きく削減される．

固有声 GMM の学習には，参照歌手と多数の事前収録目標歌手によるパラレルデータセットが必要となるが，その収録は容易ではない．そこで本稿では，VocaListener を利用して参照歌手の歌声を人工的に合成することにより，パラレルデータセットを生成する手法 [7] を提案する．本学習データ生成法では，まず，多数の事前収録目標歌手の歌声を準備する．そして，それぞれに対して VocaListener を用いて合成歌声を生成し，合成歌声と事前収録目標歌手の歌声のペアをパラレルデータセットとする．本手法では，事前収録目標歌手の歌声さえあれば，VocaListener で同一の歌声合成音源（歌声ライブラリ）を用いて模倣

*Singing voice conversion based on many-to-many eigenvoice conversion and training data generation with singing synthesis. by H. Doi, T. Toda (NAIST), T. Nakano, M. Goto (AIST) and S. Nakamura (NAIST)

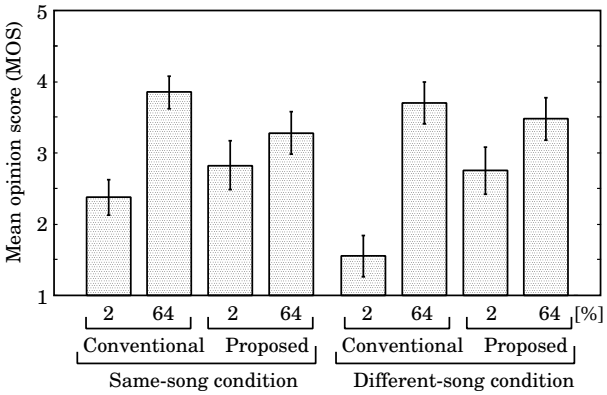

Fig. 1　音質に関する主観評価結果

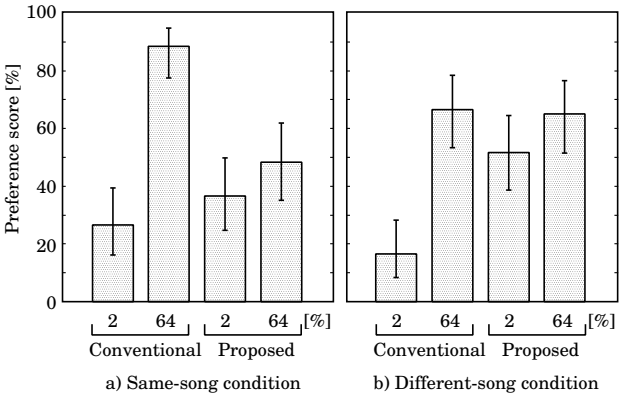

Fig. 2　話者性に関する主観評価結果

することで，参照歌手の歌声を全ての楽曲において同じ声質で用意可能である．これにより，人間の参照歌手の歌声が収録不要なので労力が少なく，曲ごとの声質の変動も抑えることができる．さらに人間とは異なり，上手く歌唱できない曲はなく，歌い回しまでも真似た歌声を生成可能なため，高品質なパラレルデータセットを構築できる．

5　実験による評価

5.1　実験条件

事前収録目標歌手の歌声として，RWC 研究用音楽データベース（ポピュラー音楽 RWC-MDB-P-2001）[8] 中の 30 曲（男性歌唱 19 曲と女性歌唱 11 曲）の無伴奏歌唱を用いる．また，参照歌手の歌声として，事前収録目標歌手の歌声を手本に VocaListener で推定された合成パラメータに基づいて，VOCALOID2（初音ミク）で合成した歌声を使用する．適応及び評価に用いる歌声として，同データベースの中から学習に未使用の 2 曲（同一歌手による RWC-MDB-P-2001 No.35 及び No.71）の無伴奏歌唱と，これら 2 曲を新たに別の女性歌手 1 名が歌唱した歌声を用いる．スペクトル特徴量として，STRAIGHT 分析 [9] により抽出された 1 次から 24 次のメルケプストラム係数を用いる．シフト長は 5 ms，サンプリング周波数は 16 kHz とする．

提案法において，スペクトル変換用固有声 GMM は，30 人（上記 RWC 研究用音楽データベースの 30 曲）の事前収録目標歌手の歌声と，それらを VocaListener で変換した参照歌手の歌声から成るパラレルデータセットから学習される．固有声 GMM の重みベクトルの次元数は 29 とし，混合数は 128 とする．比較対象として，従来の統計的手法に基づく歌声声質変換 [1] を用いる．従来法におけるスペクトル変換用 GMM の学習には，提案法における固有声 GMM の適応データとして用いる源歌手及び目標歌手の歌声と同一のものを，パラレルデータとして用いる．従来法における GMM の混合数は，評価データに対する変換精度が最大になるように，事後的に最適化する．従来法の学習データ及び提案法の適応データとして，1 曲 (RWC-MDB-P-2001 No.35) に含まれる歌声中の 2%または，64% を用い，残りの 36%を評価データとする．本稿では，主観評価を以下の 2 つの条件下で行う．

1. same-song condition: 学習・適応で用いた曲と同一の曲 (RWC-MDB-P-2001 No.35) を評価データとして使用する．
2. different-song condition: 学習・適応で用いた曲と同一歌手ではあるが異なる曲 (RWC-MDB-P-2001 No.71) を評価データとして使用する．

5.2　主観評価結果

主観評価では，各条件・手法・データ量における変換音声 (計 8 種類) の音質及び話者性を評価する．音質の評価は 5 段階の平均オピニオン評定で，話者性の評価は XAB 法で行う．被験者は共に 5 名である．尚，話者性の評価では，same-song condition と different-song condition で，目標とする曲が異なるため，それぞれ独立に評価する．

図 1 に音質の評価結果を示す．2%のデータを用いた際，different-song condition における従来法は，same-song condition の場合より，音質を大きく劣化させている．この事から，同一歌手においても，曲が異なる場合には，局所的にその声質が大きく変動することが窺える．一方，提案法は，両条件下において同等の性能を示しており，曲毎の声質の変換に対し頑健であると言える．また，少量のデータの場合は，どちらの条件下においても，提案法が従来法を上回る音質を示しており，データ量に対して頑健であることが分かる．適応データとしてパラレルデータを必要としないことも，提案法の重要な利点の一つである．

また，図 2 に話者性の評価結果を示す．話者性の評価においても，音質評価と同様の傾向を確認できる．

6　まとめ

本稿では，混合音ではない無伴奏の独唱において，任意のユーザの歌声の声質を様々な歌手の声質に自動変換できる歌声声質変換手法を提案した．また，その学習データ生成を容易にするため，歌声合成を用いた学習データ生成を提案した．実験結果から，提案法は少量かつ任意の適応データで高精度な変換が可能であることが分かった．

謝辞　本研究の一部は，科研費補助金若手研究 (A) と科学技術振興機構 OngaCREST プロジェクトによる支援を受けた．STRAIGHT の使用を許可していただいた和歌山大学河原英紀教授に感謝いたします．

参考文献

[1] 川上　他，信学技法，SP 110–297, pp. 71–76, Nov. 2010
[2] Y. Ohtani, *et al.*, Proc. INTERSPEECH, pp. 1623–1626, Sept. 2009
[3] T. Nakano *et al.*, Proc. SMC 2009, pp. 343–348, July 2009
[4] T. Toda *et al.*, IEEE Trans. ASLP, Vol. 15, No. 8, pp. 2222–2235, 2007
[5] 剣持　他，情報処理学会研究報告　音楽情報科学, Vol. 2008–MUS–74, No. 12, pp.51–56, Feb. 2008
[6] T. Toda *et al.*, Proc. ICASSP, pp. 1249–1252, Apr. 2007
[7] 土井　他，情報処理学会研究報告　音楽情報科学, Vol. 2012–MUS–96, No. 5, pp.1–9, Aug. 2012
[8] 後藤　他，情報処理学会論文誌，Vol. 45, No. 3, pp. 728–738, Mar. 2004
[9] H. Kawahara *et al.*, Speech Communication, Vol. 27, Issues. 3–4, pp. 187–207, 1999