重回帰混合正規分布モデルに基づく声質制御における精度改善*

☆久保 和隆, 小林 和弘, 戸田 智基, ニュービッグ グラム, サクティ サクリアニ,
中村 哲 (奈良先端大)

1 はじめに

声質表現語に沿った声質制御を実現する手法として，重回帰混合正規分布モデル（multiple regression Gaussian mixture model: MR-GMM）に基づく声質制御法が提案されている [1]．この手法は，声質表現語による直感的な声質制御を行うものであるが，得られる制御性能は，使用する声質表現語間の相関関係や，学習時に用いる知覚スコアの精度に大きく依存する．結果，使用者の意に沿った声質制御を十分な精度で実現できているとは言い難く，個々の要因に絞った詳細な調査が必要である．

本稿では，複数の声質表現語を用いた高精度な声質制御の実現に向けて，MR-GMM の学習に用いる知覚スコアの設計法について検討する．知覚スコア間の独立性のみでなく，対応する声質制御ベクトル間の独立性も考慮して知覚スコアを設計することで，声質制御精度を改善できることを示す．

2 修正 MR-GMM に基づく声質制御

声質表現語に基づく声質制御手法の 1 つとして，入力話者の個人性を保ちながら声質制御を実現するために，修正 MR-GMM に基づく声質制御法が提案されている [1]．修正 MR-GMM に基づく声質制御は，学習処理と変換処理で構成される．学習処理では，1 人の入力話者と S 人の事前収録目標話者が同一文セットを発話したパラレルデータを用いて，次式の MR-GMM を学習する．

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

ここで，$\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ 及び $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}]^\top$ は，入力話者と s 番目の事前収録目標話者の静的・動的特徴量ベクトルを表す．$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトル $\boldsymbol{\mu}$ 及び共分散行列 $\boldsymbol{\Sigma}$ を持つ正規分布を表す．MR-GMM の混合数は M であり，m は分布番号を示す．s 番目の事前収録目標話者に対する平均ベクトルは，次式で与えられる．

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{B}_m^{(Y)} \boldsymbol{w}(s) + \overline{\boldsymbol{\mu}}_m^{(Y)}, \quad (2)$$

ここで，$\boldsymbol{B}_m^{(Y)}$ は代表ベクトル，$\overline{\boldsymbol{\mu}}_m^{(Y)}$ はバイアスベクトルを表す．$\boldsymbol{w}(s)$ は，s 番目の話者に対する知覚スコアである．

変換処理では，声質制御対象とする目標話者の平均ベクトル $\hat{\boldsymbol{\mu}}_m^{(Y)}$ を用いて，出力平均ベクトルの表現形式を次式のように修正する．

$$\boldsymbol{\mu}_m^{(Y)}(s) = \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{B}_m^{(Y)} \Delta\boldsymbol{w}, \quad (3)$$

ここで，$\Delta\boldsymbol{w}$ は差分知覚スコアである．結果得られる修正 MR-GMM に基づき，入力話者の声質を，差分知覚スコアで表現される声質へと最尤系列変換法 [2] により変換する．

3 声質制御パラメータの学習

修正 MR-GMM による声質制御性能は，学習時に用いる知覚スコアの精度に大きく依存する．本稿では，声質表現語として，歌声声質制御において高い有効性が確認されている「知覚年齢」[1] と，知覚年齢に対して独立性が高く，かつ，知覚年齢では上手く表現できない声質を効率的に捉えることができる「通りの良さ」[3] の 2 つを用いる．1 つ目の知覚スコア（知覚年齢）が与えられているという条件の下で，2 つ目の知覚スコア（通りの良さ）を如何に設計するかという問題について，詳細に調査する．

3.1 声質制御の影響を考慮した知覚スコア付与

MR-GMM に基づく声質制御法では，音声から分析されるスペクトル特徴量や非周期成分などの分節的特徴が MR-GMM によりモデル化される．一方で，F_0 パターンやパワーパターンといった韻律的特徴は，MR-GMM によるモデル化は行わず，入力話者のものがそのまま用いられる．そのため，学習に用いる事前収録目標話者に対して知覚スコアを付与する際に，自然音声を用いると，MR-GMM に基づく声質制御法で対象外となる音響特徴量の影響を強く受けたスコアとなり，結果得られる声質制御性能の低下を引き起こす．

この問題を解決するために，本稿では，入力話者の音声から各事前収録目標話者の音声への声質変換 [2] を行い，変換音声に対して知覚スコアを付与する手法 [4] を用いる．これにより，MR-GMM による声質制御法でモデル化される音響特徴量に着目して，各事前収録目標話者に対する知覚スコアを付与することができる．この手法の有効性については，文献 [3] において報告されている．

3.2 音響空間上での独立性を考慮した知覚スコア付与

高い声質制御性能を実現するためには，互いに独立性の高い知覚スコアを用いることが重要である．また，各知覚スコアを変化させた際に生じる音響特徴量の変化成分に関しても，互いに独立であることが望まれる．文献 [3] で報告されている通り，知覚年齢と通りの良さを声質表現語として用いることで，事前収録目標話者に対して，独立性の高い知覚スコアの付与は可能である．一方で，仮に知覚スコア間で独立性が高くても，音響空間上において対応する代表ベクトル間の独立性が保障される訳ではない．

本稿では，知覚スコア間の独立性のみでなく，対応する代表ベクトル間の独立性も考慮に入れた知覚スコア付与を行い，その有効性を明らかにする．まず，知覚年齢を付与し，知覚年齢を制御可能とする修正

*Accuracy Improvement to Voice Quality Control based on Multiple-Regression Gaussian Mixture Model, by KUBO, Kazutaka, KOBAYASHI, Kazuhiro, TODA, Tomoki, NEUBIG, Graham, SAKTI, Sakriani, NAKAMURA, Satoshi (NAIST)

MR-GMM を構築する．次に，各事前収録目標話者の知覚年齢が同一となるように，修正 MR-GMM を用いて変換音声（知覚年齢正規化音声）を作成する．正規化時に用いる差分知覚年齢は，1) 音響空間上での正規化平均ベクトル間の距離最小化により決定した後に，2) 得られる知覚年齢正規化音声を聴取して知覚年齢を付与し，3) 目標とする知覚年齢との差分をさらに補正するように修正を加える，という手順により決定する [3]．これにより，各事前収録目標話者に対して，知覚年齢がほぼ同一となる知覚年齢正規化音声の作成が可能となる．これは，各事前収録目標話者の音声から，知覚年齢に対応する代表ベクトルにより表現される音響特徴量変化成分を取り除く処理とみなすことができる．この知覚年齢正規化音声に対して，通りの良さに関する知覚スコアを付与する．知覚年齢に対応する代表ベクトルと直交した音響特徴量変化成分に着目して，通りの良さに関する知覚スコアを付与することが可能となるため，結果得られる代表ベクトル間の独立性が高まることが期待される．

4　実験的評価

4.1　実験条件

MR-GMM の学習データとして，参照話者 1 名と JNAS[5] に含まれる事前収録目標話者 277 名（男性 137 名，女性 140 名）を用いる．各事前収録目標話者の発話数は，ATR 音素バランス文の 50 文程度である．サンプリング周波数は 16 kHz である．スペクトル特徴量として，STRAIGHT 分析により抽出されたスペクトル包絡から得られる 1 次から 24 次のメルケプストラム係数を用いる．音源特徴量として，F_0 と，0〜1，1〜2，2〜4，4〜6，6〜8 kHz の 5 周波数帯域において平均された非周期成分を用いる．フレームシフト長は 5 ms である．MR-GMM の混合数は，スペクトル特徴量と非周期成分でそれぞれ 256，64 である．

知覚年齢と声の通りの良さに対する差分知覚スコアの同時制御により，制御性に関する評価を行う．手法 1（3.1 節）および手法 2（3.2 節）により，各々知覚スコアを付与し，修正 MR-GMM を構築する．差分知覚スコアを，各々独立に，知覚年齢に対しては-20，0，20，声の通りの良さに対しては-10，0，10 と設定した際の声質制御音声を作成する．被験者に対して，声質制御音声をランダムな順番で提示し，知覚スコアの付与を行う．声質制御時に設定する目標話者は 10 名であり，評価話者 1 名につき，評価音声サンプル数は 24 である．なお，本稿では，被験者間における知覚スコアのばらつきの影響については調査しない．そのため，知覚スコア付与を行う被験者は，20 代男性 1 名に限定する．

4.2　実験結果

図 1 に，2 つの声質表現語を用いた同時制御に関する評価結果を示す．縦軸は，知覚年齢を表し，横軸は通りの良さを表す．黒点は，声質制御に用いる知覚スコアの設定値であり，他の点は声質制御音声に対して付与された知覚スコアである．また，表 1 に，設定した知覚スコアと評価スコア間の値の誤差を示す．表 2 には，知覚空間および音響空間における独立性に関する評価結果を示す．図 1 および表 1 の結果から，手法 2（正規化音声を用いた知覚スコア付与）は

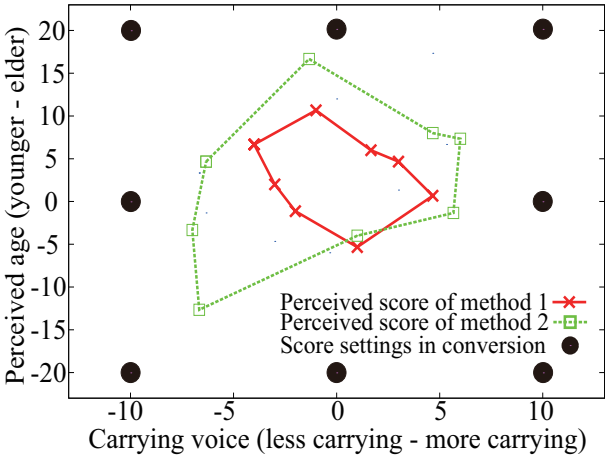

Fig. 1　2 つの声質表現語による同時制御

Table 1　目標スコアと評価スコアとの誤差

手法	通りの良さ スコアの誤差	知覚年齢 スコアの誤差
手法 1	5.46	12.0
手法 2	3.25	10.4

Table 2　知覚空間における独立性と音響空間における独立性の評価

手法	知覚スコア間の相関値	代表ベクトル間の交差角度
手法 1	-0.33	80.2 °
手法 2	-0.21	87.8 °

手法 1 と比較し，より高精度な声質制御が実現できていることが分かる．また，表 2 の結果から，手法 2 により，知覚スコア間と代表ベクトル間の双方の独立性を高めることが可能であることが分かる．

5　おわりに

本稿では，複数の声質表現語スコアによる声質同時制御において，制御性能を向上させるための知覚スコア付与手法について調査した．知覚年齢と通りの良さを対象とした実験的評価結果より，知覚年齢正規化音声に対して通りの良さの知覚スコアを付与することで，知覚スコア間の独立性と，対応する音響変化成分間の独立性の両者を満たす声質制御モデルを構築できることが分かった．今後，異なる被験者間における知覚スコアのばらつきを補正する処理に関する検討を行う．

謝辞　本研究の一部は，JSPS 科研費 26280060 および 26280061 の助成を受け実施したものである．

参考文献

[1] K. Kobayashi *et al.*, IEICE Trans. Inf. & Syst., vol.E97-D, no.6, pp.1419-1428, 2014.
[2] T. Toda *et al.*, IEEE Trans. ASLP, vol.15, no.8, pp.2222-2235, 2007.
[3] 久保和隆 他., 信学技報, vol.114, no.303, pp.65-70, 2014.
[4] K. Yamamoto *et al.*, ICASSP pp.4497-4500, 2012
[5] K. Itou *et al.*, Journal of the Acoustical Society of Japan, no.3, pp.199-206, 1995.