社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

重回帰混合正規分布モデルに基づく声質制御における制御パラメータの設計

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†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630–0192 奈良県生駒市高山町 8916 番地の 5
E-mail: †{kubo.kazutaka.kb7,kazuhiro-k,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

あらまし　本報告では，統計的手法に基づく声質制御のための制御パラメータ設計法について述べる．直感的な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）を用いて，声質表現語に対応する知覚スコアに基づき，変換音声の声質を手動制御する枠組みが提案されている．この枠組みでは，学習データとして使用する多数の事前収録目標話者の各々に対して，特定の声質表現語に基づく知覚スコアを付与することで，声質表現語と音響特徴量間の対応関係が統計的にモデル化される．そのため，得られる声質制御性能は，使用する声質表現語および付与される知覚スコアの精度に大きく依存する．本報告では，複数の声質表現語を用いた高精度な声質制御の実現を目指し，知覚スコア間における独立性と，学習された MR-GMM における各知覚スコアに対応する音響特徴量間の独立性が高まるように，声質表現語の選定および知覚スコアの付与を行う手法を提案する．実験的評価結果から，提案法により声質制御性能が改善されることを示す．
キーワード　声質制御，声質表現語，知覚スコア，重回帰混合正規分布モデル，独立性

Design of Control Parameters for Voice Quality Control based on Multiple-Regression Gaussian Mixture Model

Kazutaka KUBO†, Kazuhiro KOBAYASHI†, Tomoki TODA†, Graham NEUBIG†,
Sakriani SAKTI†, and Satoshi NAKAMURA†

† Information Science, Nara Institute of Science and Technology
Takayama 8916–5, Ikoma, Nara, 630–0192 Japan
E-mail: †{kubo.kazutaka.kb7,kazuhiro-k,tomoki,neubig,ssakti,s-nakamura}@is.naist.jp

Abstract　This report presents a method for designing control parameters in statistical voice quality control. As a method for intuitively controlling converted voice quality, a voice quality control method based on multiple-regression Gaussian mixture model (MR-GMM) was proposed. In this method, perceptual scores corresponding to voice quality expression words are manually assigned to individual pre-stored target speakers for statistically modeling a correspondence between acoustic features and the perceptual scores. Therefore, performance of voice quality control highly depends on a selection of voice quality expression words and accuracy of the perceptual score assignment. In this report, to achieve better voice quality control using multiple voice quality expression words, we propose a method for selecting voice quality expression words and assigning corresponding perceptual scores considering both independency of the perceptual scores and independency of acoustic features corresponding to them in the MR-GMM. Experimental results show that the performance of voice quality control is significantly improved by the proposed method.
Key words　voice quality control, voice quality expression words, perceptual scores, multiple regression Gaussian mixture model, independency

1. はじめに

音声合成の技術は，カーナビゲーション，Siri や Vocaloid [1] など，様々なものに使用されている．一般的に，これらのシステムで用いられている合成音声の声質は，プリセットに内蔵された声質のみに限定される．仮に，ユーザが所望する声質を持つ音声合成システムを容易に構築する事ができれば，ユーザのシステムに対する満足度が上がることが期待できる．例えば，カーナビゲーションシステムの音声を，各ユーザが自分の好みに応じて自由にセッティングできるようになると予想される．

音声合成技術の一つである声質変換においても，所望の声質の実現についての検討がなされている．声質変換は，テキスト音声合成とは異なり，音声を入力として取り扱うことができるため，リアルタイム変換が可能であり，人対人の音声コミュニケーションへの応用が期待されている [2]．特に，統計的手法に基づく声質変換（Voice Conversion）[3], [4] に関しては，歌唱音声の創作支援 [5], [6], [7] や，事故や病気などで身体的障害を患い発声機能を失ってしまった人のための発声補助など [8], [9], 実用性の高い技術の研究がなされている．これらの応用技術においては，従来の統計的声質変換で行われていた特定の話者の音声への変換のみでなく，ユーザが所望する声質へと自在に変換することが重要となる．そのため，変換音声の直感的な声質制御を可能とする技術が必要である．

直感的な声質制御を実現する技術として，任意の声質表現語に基づき目標話者の声質を制御する重回帰混合正規分布モデル（MR-GMM: multiple regression Gaussian mixture model）に基づく声質制御法が提案されている [7], [10]．変換音声の声質を声質表現語（例えば，“ 若い—老い ”や，“ 澄んだ—濁った ”など）に関する知覚スコアにより表現することで，変換音声の声質の手動制御が実現される．この技術を用いて高い声質制御性能を実現するためには，使用する声質表現語の選定および声質表現語に対応する知覚スコアの設計が重要となる．例えば，変換音声の声質との対応が不十分な知覚スコアを用いても，高い制御性能は得られない．また，個々の声質表現語に対する知覚スコアの精度が十分に高い場合においても，お互いに強く依存している場合には，制御性能が劣化すると考えられる．

本報告では，MR-GMM に基づく声質制御において，操作性に優れた声質表現語の選定および知覚スコアの設計に取り組む．知覚スコア間の独立性と，各知覚スコアに対応する声質成分間の音響空間上での独立性に着目し，声質表現語の選定を行う．また，声質制御により得られる変換音声に特化した知覚スコアの設計を行う．実験的評価結果から，提案手法を用いることで，MR-GMM に基づく声質制御性能を大幅に改善できることを示す．

2. 関連研究

学習データに含まれない任意の入力話者と任意の目標話者の間で声質変換を可能とするために，固有声変換技術 [11], [12] が提案されている．学習処理として，一人の参照話者と多数の事前収録目標話者のパラレルデータセットを用いて，固有声 GMM（EV-GMM: Eigenvoice GMM）を学習する．任意の入力話者および目標話者に対する変換モデルは，各話者の極少量かつ任意の音声データを用いて，EV-GMM の適応パラメータを各々独立に推定することで，容易に構築することができる．また，音声データが用意出来ない場合においても，EV-GMM の適応パラメータを手動操作することで，変換音声の声質を変化させることが可能である．EV-GMM において，適応パラメータの個々の次元が表す声質成分は，音響空間上で直交しているため，複数の適応パラメータを同時に操作する際には，操作性の面で利点がある．しかしながら，適応パラメータの空間は人の主観に沿っていないため，各次元で表される声質成分を具体的に説明することは困難である．パラメータ操作に伴う変換音声の声質を予測すること自体が困難であるため，所望の声質の実現には，パラメータ変更と変換音声の確認といった手順を何度も試行する必要がある．

一方で，直感的な声質制御を実現する技術として，統計的パラメトリック音声合成の研究において，容易に解釈可能な声質制御パラメータを用いた声質制御手法が提案されている．隠れマルコフモデル（HMM: Hidden Markov Model）に基づくテキスト音声合成 [13] において，発話様式を表す低次元ベクトルから HMM の平均ベクトルへの写象を表現する重回帰 HMM を用いることで，合成音声の発話様式の手動制御が実現されている [14]．さらに，“ 暖かい—冷たい ”などの声質表現語に対する主観評価値で構成される低次元ベクトルを導入することで，声質表現語に基づく合成音声の声質制御も実現されている [15]．本枠組みは，声質変換にも応用されており，MR-GMM に基づく声質制御においてその有効性が確認されている [10]．

3. 統計的手法に基づく声質変換・制御

3.1 GMM に基づく声質変換

統計的手法に基づく声質変換は，ある特定話者（入力話者）の声質を異なる特定話者（目標話者）の声質へと変換する技術である．代表的な手法である GMM に基づく声質変換では，主に分節的特徴（スペクトル包絡等）の変換を対象としており，リアルタイム変換処理も実現可能である [2]．本手法は，学習処理と変換処理で構成される．

学習処理では，入力話者と目標話者のパラレルデータより音響特徴量を抽出し，GMM によりそれらの結合確率密度関数をモデル化する．入力話者と目標話者の音響特徴量を，$2D$ 次元の結合静的・動的特徴量ベクトル $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ とする．ここで，$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ は，フレーム t における入力話者と目標話者の静的音響特徴量であり，$\Delta\boldsymbol{x}_t$ と $\Delta\boldsymbol{y}_t$ は，同フレームの入力話者と目標話者の動的特徴量である．$\top$ は転置を表す．これらの音響特徴量の結合確率密度関数は，以下の式により与えられる．

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

ここで $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ の正規

分布を表す．混合数は M であり，m は分布番号を表す．$\boldsymbol{\lambda}$ は GMM のパラメータセットを表し，個々の分布における分布重み α_m，平均ベクトル $\boldsymbol{\mu}_m$，共分散行列 $\boldsymbol{\Sigma}_m$ を含む．パラレルデータに対して，動的時間伸縮により対応づけられた $\boldsymbol{X}_t$，$\boldsymbol{Y}_t$ を用いて GMM を学習する．

変換処理では，入力話者の音声から抽出された音響特徴量を最尤系列変換法 [4] により目標話者の音響特徴量へと変換する．入力話者と目標話者の特徴量系列ベクトルを，$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ と $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ とする．ここで，T はフレーム数である．変換された静的特徴量系列 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ は次式で示される．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{2}$$

ここで $\boldsymbol{W}$ は静的特徴量系列を結合静的・動的特徴量系列に拡張する行列である．条件付き確率密度関数 $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ は，式（1）で与えられた結合確率密度関数から解析的に導出される．なお，過剰な平滑化による変換音声の音質劣化を緩和するため，系列内変動（GV: Global Variance）[4] を考慮する．

3.2 MR-GMM に基づく声質制御

主観的な声質表現語に基づく声質制御を実現するため，MR-GMM に基づく声質制御を行う．GMM に基づく声質変換と同様に，学習処理と変換処理により構成される．

学習処理では，一人の参照話者と複数の事前収録目標話者の音響特徴量の結合確率密度関数を学習する．MR-GMM による結合確率密度関数は以下のように与えられる．

$$\begin{aligned} &P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(MR)}, \boldsymbol{w}(s)\right) \\ &= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(s)\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right). \end{aligned} \tag{3}$$

s 番目の事前収録目標話者の平均ベクトルは以下の式で与えられる．

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{B}_m^{(Y)}\boldsymbol{w}(s) + \overline{\boldsymbol{\mu}}_m^{(Y)}, \tag{4}$$

ここで，$\boldsymbol{B}_m^{(Y)}$ および $\overline{\boldsymbol{\mu}}_m^{(Y)}$ は，それぞれ代表ベクトルとバイアスベクトルを表す．$\boldsymbol{w}(s)$ は，s 番目の事前収録目標話者の音声に対して聴取実験により得られた各声質表現語に対する知覚スコアからなるベクトルを表す．

変換処理において，MR-GMM における出力側の平均ベクトルは，各声質表現語スコアに対する所望の知覚スコアを入力することで決定される．入力話者の音声は，GMM に基づく声質変換と同様の処理で変換される．

3.3 声質表現語に基づく個人性を保持した声質制御

個人性を保持した声質表現語に基づく声質制御のために，式（4）における出力平均ベクトルの表現方法を変更する．式（4）で表現される出力平均ベクトルは，事前収録目標話者の平均的な声質を基準として，各声質表現語の知覚スコアによる変動を捉えるものであるため，特定話者の声質を表現するものではない．これに対し，入力話者の声質を基準とした声質制御を実現するために，入力話者の個人性を保持した声質制御法が提案されている．[7] この手法では，MR-GMM における出力平均ベクトルの表現方法を以下のように変更する．

$$\begin{aligned} \boldsymbol{\mu}_m^{(Y)}(s) &= \boldsymbol{B}_m^{(Y)}\boldsymbol{w}(s) + \overline{\boldsymbol{\mu}}_m^{(Y)} \\ &= \boldsymbol{B}_m^{(Y)}(\boldsymbol{w}^{(i)} + \Delta\boldsymbol{w}(s)) + \overline{\boldsymbol{\mu}}_m^{(Y)} \\ &= \boldsymbol{B}_m^{(Y)}\boldsymbol{w}^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{B}_m^{(Y)}\Delta\boldsymbol{w}(s) \\ &\simeq \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{B}_m^{(Y)}\Delta\boldsymbol{w}(s), \end{aligned} \tag{5}$$

ここで，$\hat{\boldsymbol{\mu}}_m^{(Y)}$ は入力話者の平均ベクトルであり，本報告ではパラレルデータによって得られる最尤推定値を用いる．各声質表現語差分知覚スコア $\Delta\boldsymbol{w}(s)$ は，入力話者の知覚スコア $\boldsymbol{w}^{(i)}$ を基準とした差分量で表される．

4. 声質制御パラメータの設計

MR-GMM における声質制御において，複数の声質表現語を用いる際には，各声質表現語に対する知覚スコア間の独立性（知覚空間における独立性）と，各声質表現語に対応する MR-GMM の代表ベクトル間の独立性（音響空間における独立性）の両者を満たすように，声質制御パラメータを設計することで，より高い制御性能が得られると予想される．本報告では，一つ目の声質表現語として，歌声声質制御においてその有効性が確認されている知覚年齢 [7] を採用した際に，上記の独立性を考慮して二つ目の声質表現語を選定する方法について調査する．

まず，MR-GMM に基づく声質制御において，声質表現語を付与するために使用する評価音声について調査する．次に，知覚年齢を正規化した音声を使用することで，知覚年齢と独立性の高い声質表現語の選定を試みる．得られた知覚スコアに基づき，MR-GMM を構築することで，二つの声質表現語に関する知覚スコアに基づく声質制御を実現する．

4.1 統計的声質制御の影響を考慮した評価音声の使用

MR-GMM の学習では，各事前収録目標話者に対する声質表現語の知覚スコアが必要である．従来では，各事前収録目標話者に対する知覚スコアは，自然音声に対して付与されることが多い [10], [16], [17]．一方で，MR-GMM では，主に分節的特徴をモデル化する枠組みであるため，韻律的特徴に対しては高度な変換処理を行わない．また，分節的特徴に限っても，自然音声の特徴が完全に再現されるわけではなく，少なからず変換誤差の影響を受ける．そのため，自然音声に対して付与された知覚スコアは，MR-GMM における声質制御において最適とは言い難い．

MR-GMM において制御可能な音響特徴量に着目した知覚スコア付与を実現するために，変換音声に対して知覚スコアの付与を行う．参照話者と事前収録目標話者のパラレルデータを用いて，参照話者から各事前収録目標話者に対する変換モデルを構築し，参照話者から各事前収録目標話者への変換音声を作成する．この変換音声は，参照話者の韻律的特徴を保持しており，F_0 に関しては平均値と標準偏差のみを変更したものとなる．また，分節的特徴は変換誤差の影響を受けたものとなる．この変換音声に対して知覚スコアを付与することで，MR-GMM で制御可能な音響特徴量のみに着目した知覚スコアが得られると期

待される．

4.2 知覚年齢正規化音声の使用

一つ目の声質表現語である知覚年齢に対する知覚スコアと比較して，知覚空間における独立性と音響空間における独立性が高くなるような声質表現語を選定するために，知覚年齢の影響を取り除いた変換音声の使用を検討する．知覚年齢が音響特徴量に与える影響を取り除くために，各事前収録目標話者に対して，共通の知覚年齢を持つ変換音声を作成する．3.3 節で述べた個人性を保持した声質制御法を応用し，各事前収録目標話者の話者性を保持した知覚年齢正規化を行う．以下では，知覚年齢正規化を行うための，差分知覚年齢スコアの設定法について述べる．

4.2.1 知覚年齢スコアを基準とした知覚年齢正規化法

変換音声の知覚年齢を基準として，知覚年齢を正規化する．s 番目の事前収録目標話者の知覚年齢を $w^{(O)}(s)$，目標とする正規化知覚年齢を $w^{(T)}$ とすると，s 番目の事前収録目標話者に対する差分知覚年齢スコア $\Delta w(s)$ は以下の式で表される．

$$\Delta w(s) = w^{(T)} - w^{(O)}(s) \tag{6}$$

4.2.2 音響空間における距離尺度を利用した知覚年齢正規化法

音響空間において，目標知覚年齢の平均ベクトル $\boldsymbol{\mu}_m^{(T)}$ と s 番目の事前収録話者の平均ベクトル間の誤差最小化基準に基づき，$\Delta w(s)$ を決定する．

$$\begin{aligned}\Delta w(s) = \arg\min_{\Delta w(s)} &(\boldsymbol{\mu}_m^{(T)} - \boldsymbol{\mu}_m^{(Y)}(s) - \boldsymbol{b}_m^{(Y)}\Delta w(s))^\top \\ &\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m^{(T)} - \boldsymbol{\mu}_m^{(Y)}(s) - \boldsymbol{b}_m^{(Y)}\Delta w(s))\end{aligned} \tag{7}$$

$$\boldsymbol{\mu}_m^{(T)} = \boldsymbol{b}_m^{(Y)} w^{(T)} + \overline{\boldsymbol{\mu}}_m^{(Y)} \tag{8}$$

4.2.3 正規化音声に対する知覚年齢を考慮した知覚年齢正規化法

知覚年齢の制御精度は十分なものではなく，上述の手法により，事前収録目標話者の知覚年齢を目標とする正規化知覚年齢へと完全に正規化する事は困難である．そこで，正規化音声に対して再度知覚年齢を付与し，その値を利用することで，正規化精度を高める．正規化時に使用した差分知覚年齢を $\Delta w^{(d)}(s)$，正規化音声に対して付与された知覚年齢を $\hat{w}(s)$ とする．差分知覚年齢が 0 の際には知覚年齢が $w^{(O)}(s)$ となることから，これらを線形に補間することで，知覚年齢が $w^{(T)}$ となるような正規化音声を生成すると予測される差分知覚年齢を次式により決定する．

$$\Delta w(s) = \Delta w^{(d)}(s) \frac{w^{(T)} - w^{(O)}(s)}{\hat{w}(s) - w^{(O)}(s)} \tag{9}$$

4.3 因子分析による声質表現語の決定

知覚年齢に対して独立性の高い声質表現語を決定するために，知覚年齢正規化音声に対して， 声質表現語対 [16], [17] の中から表 1 に示す了解性と同義性の高い 14 語に基づき，知覚スコアの付与を行う．知覚年齢正規化音声の使用により，知覚年齢では表現できない音響特徴量に着目した知覚スコア付与がなされると期待される．得られた 14 種類の知覚スコアに対し，因子分析を行うことで，二つ目の声質表現語を決定する．

表 1　使用した声質表現語

太い声	かすれた声	張りのある声
通りの良い声	だみ声	響きのある声
細い声	鼻にかかった声	がらがら声
澄んだ声	弱々しい声	濁った声
張りのない声	潰れた声	

5. 実験的評価

まず予備実験として，評価音声に対して付与される知覚スコアの安定性を確認する．次に，知覚年齢正規化法ついて評価する．そして，知覚年齢正規化音声に対して各種声質表現語に対する知覚スコアを付与し，因子分析を行うことで，知覚年齢との併用を考えた際に最適と思われる声質表現語の選定を行う．最後に，知覚年齢と選定された声質表現語を用いて，声質の同時制御を行い，その制御性能を評価する．

5.1 実験条件

JNAS [18] に含まれる 20，30，40，50，60 歳代の日本人男女約 300 名分の音声データを用いる．個々の話者は，ATR 音素バランス文 50 文程度を発声している．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 [19] 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 の学習データとして，参照話者 1 名と事前収録目標話者 277 名（男性 137 名，女性 140 名）を用いる．MR-GMM の混合数は，スペクトル特徴量に対しては 256，非周期成分に対して 64 とする．

5.2 知覚スコア付与の安定性に関する評価

手動で付与される知覚スコアの安定性を評価するために，同一音声に対して複数回知覚スコアを付与した際に，知覚スコアの平均値が安定する回数を調査する．実験には，事前収録目標話者の中からランダムに選択した 30 名の話者の音声 1 文を用いる．計 30 種類の音声をランダムに提示し，知覚年齢を付与する．各音声に対しては，計 12 回知覚年齢の付与を行う．被験者は 20 代男性 1 名である．

図 1 に知覚年齢付与の安定性に関する評価結果を示す．横軸は 1 つの音声サンプルに対し知覚年齢を評価した回数，縦軸は得られる平均知覚年齢同士の相関係数の平均を表す．例えば，評価回数が 3 の場合は，12 回の評価結果からランダムに選択された 3 回分を用いて平均知覚年齢を計算し，この処理を繰り返すことで，平均知覚年齢間の相関係数を求めた結果を示す．評価回数の増加に伴い相関係数が上昇することから，同一音声に対して複数回知覚スコア付与を行うことで，より安定した知覚

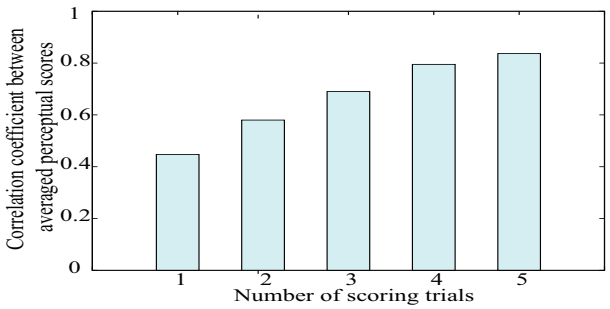

図 1　知覚年齢スコアの安定性に関する評価

表 2　各処理で得られる変換音声に対する知覚年齢と基準となる知覚年齢間の相関係数（1 回目の実験）

	相関係数
年齢正規化なし	0.86
知覚年齢基準の正規化	0.81
音響的距離尺度基準の正規化	0.78

スコアが得られることが分かる．とりわけ，5 回評価した際には，相関係数が 0.8 を上回ることが分かる．

5.3　5．2 知覚年齢正規化法の評価

MR-GMM による知覚年齢正規化精度を評価する．先の実験で用いた評価話者 30 名の音声 1 文に対して，知覚年齢変換をしない変換音声（差分知覚年齢を 0 に設定した際の変換音声）と知覚年齢正規化音声を作成する．知覚年齢正規化においては，全事前収録目標話者の平均知覚年齢である 45 歳を目標値とする．知覚年齢正規化手法として，まず，4.2.1 節で述べた知覚年齢を基準とした手法と，4.2.2 節で述べた音響空間における距離尺度を利用した手法を用いる．これらの音声をランダムな順番で提示して，知覚年齢付与を行う．被験者は先の実験と同様の 20 代男性 1 名であり，同一の音声に対して 5 回知覚年齢付与を行い，平均知覚年齢を求める．また，参照話者から各評価話者への変換音声に対しても知覚年齢を付与し，得られた結果から各評価話者の基準となる知覚年齢を求める．

表 2 に，基準となる知覚年齢に対する各種知覚年齢の相関係数を表す．知覚年齢正規化を施すことで，相関係数が下がることから，その効果が確認できる．また，知覚年齢基準の手法と比較し，音響空間における距離尺度を利用した手法の方が，より効果的に知覚年齢を正規化できることが分かる．一方で，未だ基準となる知覚年齢に対し，高い相関が見られており，十分な正規化精度が得られているとは言い難い．

次に，音響空間における距離尺度を利用した正規化法で得られた知覚年齢を用いて，4.2.3 節で述べた手法により知覚年齢正規化音声を作成する．先の実験と同様の条件で，知覚年齢変換をしない変換音声と，音響空間における距離尺度を利用した手法による正規化音声，さらに正規化音声に対する知覚年齢を利用した手法による正規化音声に対して，知覚年齢の付与を行う．

表 3 に結果を示す．正規化音声に対する知覚年齢を利用した手法を用いることで，相関係数を大幅に下げることができる．このことから，本手法が最も高い精度で知覚年齢正規化を行えることが分かる．

表 3　各処理で得られる変換音声に対する知覚年齢と基準となる知覚年齢間の相関係数（2 回目の実験）

	相関係数
年齢正規化なし	0.80
音響的距離尺度基準の正規化	0.67
上記の知覚年齢を考慮した正規化	0.34

表 4　因子分析結果

項目	第一因子	第二因子	第三因子
通りの良い声	**-.780**	-.157	.035
響きのある声	**-.707**	.267	-.141
張りのある声	**-.705**	-.078	.168
澄んだ声	**-.582**	-.115	-.240
だみ声	-.182	**.780**	-.092
太い声	-.088	.166	**.723**
細い声	-.068	.215	**-.763**
がらがら声	-.017	**.470**	-.032
鼻にかかった声	.121	.140	.194
つぶれた声	.245	.429	.159
濁った声	**.455**	.185	.100
弱々しい声	**.511**	-.019	-.426
かすれた声	**.589**	-.120	-.073
張りのない声	**.712**	-.049	-.002
因子寄与	3.633	2.024	1.566

5.4　知覚年齢と独立な声質表現語の抽出

事前収録目標話者 277 名に対して，4.2.3 節で述べた手法により知覚年齢正規化音声を作成し，表 1 に示す各声質表現語に関する 7 段階知覚スコア（-3〜3）の付与を行う．用いる音声は各事前収録目標話者において 1 文であり，被験者は先の実験と同じ 20 代男性 1 名である．得られた知覚スコアに対して因子分析（最尤法とプロマックス回転を使用）を行う．

表 4 に，各声質表現語に対する因子分析の結果を示す．第一因子は，声の通りの良さや響き，張りといった要因を捉えていることが分かる．この結果を踏まえて，知覚年齢の次に使用する声質表現語として,「声の通りの良さ」を使用する．

5.5　複数の声質表現語による同時制御に対する評価

「知覚年齢」と「声の通りの良さ」といった二つの声質表現語に対する知覚スコアを同時に用いて，声質制御を行い，その制御性能を評価する．MR-GMM の学習に用いる知覚スコアとして，手法 1）自然音声に対して付与された知覚スコア，手法 2）変換音声に対して付与された知覚スコア，手法 3）変換音声と年齢正規化音声に対して付与された知覚スコアの 3 種類を用いる．知覚スコアの付与は，これまでの実験と同様の手順で，20 代男性 1 名により行う．各種知覚スコアを用いて学習される 3 種類の MR-GMM を用いて，知覚年齢と声の通りの良さに関する差分知覚スコアを（20，-10），(20，0），(20，10），(0，-10），(0，10），(-20，-10），(-20，0），(-20，10）と変化させることで，計 24 通りの変換音声を作成する．これら 24 通りの変換音声を，同 20 代男性 1 名の被験者にランダムな順番で提示し，被験者は「知覚年齢」と「声の通りの良さ」に関する差分知覚スコアを付与する．

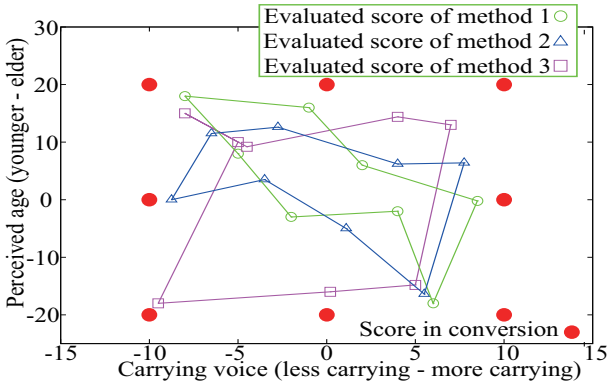

図 2　複数の声質表現語による同時制御

表 5　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	-0.29	57.1 °
手法 2	-0.33	80.2 °
手法 3	-0.21	87.8 °

図 2 に評価結果を示す．縦軸が「知覚年齢」，横軸が「声の通りの良さ」に関する差分知覚スコアを表している．手法 1 および手法 2 の知覚スコアを用いた際には，二つの知覚スコア間において相関が強く出ていることから，両スコアを独立に操作することが困難であることがわかる．一方で，手法 3 の知覚スコアを用いた際には，二つの知覚スコア間の独立性を保ちつつ，比較的高い精度で声質制御が可能であることが分かる．また，各手法で得られる知覚スコア間の相関係数の平均値と，学習された MR-GMM の代表ベクトル間の交差角度 (cos 距離を角度で表したもの) を表 5 に示す．手法 3 を用いることで，知覚スコア間の相関を低く抑えることができ，かつ，ほぼ直交する代表ベクトルが得られることが分かる．このことから，手法 3 により，知覚空間の独立性と音響空間における独立性を両立させる知覚スコア設計が可能であることが分かる．

6. 結　　論

本報告では，統計的手法に基づく声質制御の為の，制御性に優れた制御パラメータの設計法について提案した．一つ目の声質表現語として知覚年齢を採用するという条件の下で，知覚スコア間の独立性と，各知覚スコアに対応する声質成分間の音響空間上での独立性の両者を考慮することで，二つ目の声質表現語の選定を行った。また，統計的手法に基づく声質制御に特化した知覚スコア付与を行った。得られた知覚スコアに基づき，重回帰混合正規分布モデルを学習し，その声質変換精度を評価した結果，提案する知覚スコア付与手法を用いることで，声質制御性能が向上する事を示した．

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．

文　　献

[1] H. Kenmochi, and H. Ohshita, “VOCALOID – Commericial singing synthesizer based on sample concatenation,” Proc. INTERSPEECH, pp.4011–4012, Aug. 2007.

[2] T. Toda, T. Muramatsu, and H. Banno, “Implementation of computationally efficient real-time voice conversion,” Proc. INTERSPEECH, Sept. 2012.

[3] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. SAP, vol.6, no.2, pp.131–142, Mar. 1998.

[4] T. Toda, A.W. Black, and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. ASLP, vol.15, no.8, pp.2222–2235, Nov. 2007.

[5] M. Morise, M. Onishi, H. Kawahara, and H. Katayose, “v.morish ’09: A morphing-based singing design interface for vocal melodies,” Proc. ICEC, pp.185–190, 2009.

[6] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, “Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system,” Proc. APSIPA ASC, Nov. 2012.

[7] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, “Voice Timbre Control Based on Perceived Age in Singing Voice Conversion,” IEICE Transactions on Information and Systems, vol.E97-D, no.6, pp.1419–1428, 2014.

[8] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, “Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion,” Audio, Speech, and Language Processing, IEEE/ACM Transactions on, vol.22, no.6, pp.172–183, 2014.

[9] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Communication, vol.54, no.1, pp.134–146, 2012.

[10] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Regression approaches to voice quality control based on one-to-many eigenvoice conversion,” 6th ISCA Speech Synthesis Workshop (SSW6), pp.101–106, Aug. 2007.

[11] T. Toda, Y. Ohtani, and K. Shikano, “One-to-many and many-to-one voice conversion based on eigenvoices,” Proc. ICASSP, pp.1249–1252, Apr. 2007.

[12] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Many-to-many eigenvoice conversion with reference voice,” Proc. INTERSPEECH, pp.1623–1626, Sept. 2009.

[13] H. Zen, K. Tokuda, and A.W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol.51, no.11, pp.1039–1064, Nov. 2009.

[14] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style control technique for HMM-based expressive speech synthesis,” IEICE Trans. Information and Systems, vol.E90-D, no.9, pp.1406–1413, Sep. 2007.

[15] M. Tachibana, T. Nose, J. Yamagishi, and T. Kobayashi, “A technique for controlling voice quality of synthetic speech using multiple regression HSMM,” Proc. INTERSPEECH, pp.2438–2441, Sept. 2006.

[16] 木戸博，粕谷英樹，“通常発話の声質に関連した日常表現語の抽出，” 日本音響学会誌，vol.55，no.6，pp.405–411，1999.

[17] 木戸博，粕谷英樹，“通常発話の声質に関連した日常表現語，” 日本音響学会誌，vol.57，no.5，pp.337–344，2001.

[18] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, and K. Shikano, “Japanese Newspaper Article Sentences,” Journal of the Acoustical Society of Japan, no.3, pp.199–206, 1995.

[19] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system straight,” Proc. MAVEBA, Sept. 2001.