Vol.2011-SLP-85 No.11
2011/2/5

統計的声質変換に基づく食道音声強調における声質制御

山本憲三[†1] 土井啓成[†1] 戸田智基[†1]
猿渡洋[†1] 鹿野清宏[†1]

喉頭摘出者のための代替発声法の一つに，食道発声法がある．食道発声法により生成される食道音声は，健常者の音声と比較すると，音質が低く，また，話者によらず似た声質となり話者性が劣化する．食道音声の品質改善のために，一対多固有声変換に基づく食道音声から健常者音声への変換 (Esophageal Speech-to-Speech: ES-to-Speech) が提案されており，その有効性が示されている．この手法では，目標とする音声データに対して，変換音声の声質を自動的に適応させることができる．一方で，目標とする音声データが手に入らない際に，所望の声質を実現させるのは容易ではない．本稿では，利用者による直感的な声質手動制御を実現するために，ES-to-Speech に対し，重回帰混合正規分布モデル (Multiple Regression Gaussian Mixture Model: MR-GMM) に基づく声質変換・制御法を導入する．また，声質制御性能を向上させるために，MR-GMM の学習に用いる声質表現語スコアの付与方法や，回帰分析手法についても検討する．

Voice Quality Control in Esophageal Speech Enhancement Using Statistical Voice Conversion

Kenzo Yamamoto,[†1] Hironori Doi,[†1] Tomoki Toda,[†1]
Hiroshi Saruwatari[†1] and Kiyohiro Shikano [†1]

Esophageal speech is one of the alternative speaking methods for total laryngectomees. Compared with normal speech, its sounds are unnatural and hard to distinguish from esophageal speech produced by other laryngectomees due to lack of speaker dependent features in voice quality. To improve naturalness and speaker individuality of the esophageal speech, a conversion method from esophageal speech to normal speech (ES-to-Speech) using one-to-many eigenvoice conversion has been proposed. This method is capable of adapting the converted voice quality to given target speech data. However, it is hard to manually control the conveted voice quality when target speech data are not available. In this report, we introduce a voice quality control method based on multiple regression Gaussian mixture model (MR-GMM) to ES-to-Speech to make it possible to intuitively control the converted voice quality. Moreover, to further improve the performance of voice quality control, we improve a method for assigning voice quality scores into individual speakers' voices and a regression analysis method.

1. はじめに

音声は，人間の最も主流なコミュニケーション手段の一つであるが．一般に発声障害者と呼ばれる者は音声の生成に何らかの障害を抱えており，音声コミュニケーションにおいて大きな困難を有する．発声障害者の中でも，事故や喉頭癌等によって喉頭を摘出された喉頭摘出者は，喉頭と共に声帯も失うため，自身の声帯振動による音源を用いた発声が不可能である．そのため，喉頭摘出者は健常者とは別の方法で音源を得る必要がある．

喉頭摘出者のための代用発声法の一つに食道発声法がある．これは，胃に空気を飲み込み，吐き出す際に食道入り口付近の粘膜のひだを振動させることにより音源を生成し，調音することで発声を行う手法である．食道発声法は音源を体内で生成するため，電気式人工喉頭のような他の手法とは異なり，生成される音声には肉声感があり，器具を使用せずに発声することが可能である．また，日本では，ボランティア団体による食道発声法の習得支援環境が充実していることもあり，食道発声法は広く使用されている．しかしながら，食道発声法で生成される食道音声は，健常者の音声と比較するとその音質は低く，また，話者によらず似た声質となり話者性が劣化する．

食道音声の音質改善のために，これまでに，統計的声質変換[1),2)] に基づく食道音声から健常者音声への変換 (Esophageal Speech-to-Speech: ES-to-Speech)[3)] が提案されており，その有効性が示されている．また，話者性改善のために，一対多固有声変換[4)] に基づく ES-to-Speech[3)] が提案されている．この手法では，目標とする音声データに対して，変換音声の声質を自動的に適応させることができる．一方で，目標とする音声データが手に入らない際に，所望の声質を実現させるのは容易ではない．

本稿では，ES-to-Speech に対し，重回帰混合正規分布モデル (Multiple Regression Gaussian Mixture Model: MR-GMM) に基づく声質変換・制御法[5)] を導入することで，操作性に優れた変換音声の声質制御を実現する．また，声質制御性能の向上のために，MR-GMM の学習に用いる声質表現語スコアの付与方法や，回帰分析手法についても検討する．

†1 奈良先端科学技術大学院大学　情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

Vol.2011-SLP-85 No.11
2011/2/5

2. 食道音声から健常者音声への変換: ES-to-Speech[3)]

ES-to-Speech における学習処理として，次式により入力特徴量 $\boldsymbol{X}_t$ と出力特徴量 $\boldsymbol{Y}_t$ の結合確率密度を Gaussian Mixture Model (GMM) でモデル化する[6)]．

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

ここで，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ の正規分布であり，α_m は m 番目の分布重み，M は GMM の混合数である．

ES-to-Speech では，入力特徴量をスペクトルセグメント特徴量[7)] とし，出力特徴量をそれぞれ，スペクトル，F_0，非周期成分とする GMM を独立に学習し，各出力特徴量の推定に用いる．また，目標音声の静的特徴量系列に対する分散として定義される系列内変動 (Global Variance: GV) ベクトル $\boldsymbol{v}(\boldsymbol{y})$ を発話ごとに求め，その確率密度分布 $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ を正規分布でモデル化する[2)]．

変換時は，入出力の特徴量系列をそれぞれ $\boldsymbol{X} = [\boldsymbol{X}_1{}^\top, \cdots, \boldsymbol{X}_T{}^\top]^\top$, $\boldsymbol{Y} = [\boldsymbol{Y}_1{}^\top, \cdots, \boldsymbol{Y}_T{}^\top]^\top$ とし，尤度関数 $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})^\omega$ の最大化に基づき，$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ に従う静的特徴量系列 $\boldsymbol{y}$ を推定する．ここで，$\boldsymbol{W}$ は静的特徴量系列を静的・動的特徴量系列へと変換する行列であり，ω は $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ と $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ の比率を制御する重みである．

3. 重回帰混合正規分布モデル (MR-GMM) に基づく声質制御法[5)]

MR-GMM に基づく声質制御法では，多数の事前収録出力話者に対して，声質の具体的な特徴を表す形容詞対（声質表現語対[8)9)]）に対応する声質表現語スコアを人手で付与する．また，個々の事前収録出力話者に対して，変換モデルを学習する．そして，事前収録出力話者に対する声質表現語スコアと変換モデルパラメータの対応関係を重回帰分析によりモデル化する．これにより，声質表現語スコアの手動操作による直感的な声質制御が可能となる．

MR-GMM において，m 番目の分布に対する出力平均ベクトルは，次式のようにバイアスベクトル $\boldsymbol{b}_m^{(0)}$ と J 個の各声質表現語対が表す声質をモデル化する代表ベクトル $\boldsymbol{B}_m = \left[\boldsymbol{b}_m^{(1)}, \boldsymbol{b}_m^{(2)}, \cdots, \boldsymbol{b}_m^{(J)}\right]$ の線形結合として表される．

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

すなわち MR-GMM は，フリーパラメータである J 次元の声質表現語スコアベクトル $\boldsymbol{w} = [w_1, w_2, \cdots, w_J]^\top$ と，全事前収録出力話者に対して共通のパラメータセット $\boldsymbol{\lambda}^{(MR)} = \{\alpha_m, \boldsymbol{\mu}_m^{(X)}, \boldsymbol{b}_m^{(0)}, \boldsymbol{B}_m, \boldsymbol{\Sigma}_m^{(X,Y)}\}$ で表され，声質表現語スコアベクトルを操作することで，出力平均ベクトルを変化させることができる．

4. ES-to-Speech における声質制御

4.1 MR-GMM に基づく ES-to-Speech 変換音声の声質制御法

食道音声の直感的な声質制御を目指し，MR-GMM に基づく声質制御法を ES-to-Speech に導入する．本手法は，学習部と変換部から成る．

学習部では，食道音声のスペクトルセグメント特徴量を入力とし，健常者のスペクトルまたは非周期成分を出力とする MR-GMM，及び，健常者の F_0 を出力とする GMM[3)] をそれぞれ独立に学習する．各 MR-GMM の学習では，各事前収録出力話者に対して声質表現語スコアベクトルと話者依存 GMM を構築し，声質表現語スコアベクトルから出力平均ベクトルへ変換する回帰行列パラメータ（式 (2) における $\{\boldsymbol{b}_m^{(0)}, \boldsymbol{B}_m\}$）を最小二乗推定する[5)]．$F_0$ 推定用 GMM の学習では，食道音声の抑揚を再現するため，食道音声の抑揚を模して発声された健常者音声を目標音声として用いる．F_0 に関しては，声質表現語スコアベクトルを説明変数，各事前収録出力話者音声の対数 F_0 の平均・標準偏差を目的変数とする重回帰分析を用いることで，声の高さや変動の大きさといった大局的な特徴のみを声質表現語スコアベクトルで制御する．

変換部では，与えられた声質表現語スコアベクトルに基づき，スペクトル及び非周期成分推定用 MR-GMM の出力平均ベクトルを求め，食道音声のスペクトルセグメント特徴量を声質表現語スコアベクトルに対応した声質を表すスペクトル及び非周期成分へと変換する．F_0 に関しては，GMM を用いて食道音声のスペクトルセグメント特徴量から F_0 パターンを推定した後に，声質表現語スコアベクトルから求められる対数 F_0 の平均・標準偏差に沿うように，推定された F_0 パターンに対して線形変換を行う．なお，本稿では，GV のモデルパラメータは，全事前収録出力話者のデータを用いて推定される話者非依存のものを用いる．

4.2 スコアリング対象音声についての検討

従来の MR-GMM に基づく声質制御法では，学習時に，事前収録出力話者の自然音声に対し，声質表現語に関するスコアリングを行う．この場合，自然音声が持つ様々な特徴がスコアリング結果に影響を与える．しかしながら，本手法において，声質表現語スコアベクトルにより操作する音響特徴量は，スペクトル，非周期成分，および F_0 の大局的な特徴のみであり，例えば，話速やイントネーションなどは操作しない．また，スペクトル変換においても，目標音声のスペクトルを完全に再現できるわけでなく，変換処理を通して失われる特徴も存在する．すなわち，本手法で制御できる声質は，変換処理で実現できるものに限定される．そのため，自然音声に対して付与されたスコアを用いて変換モデルを構築すると，利用者の意図する声質と変換音声の声質との間に差が生じると考えられる．

そこで本稿では，従来の ES-to-Speech により食道音声から各事前収録出力話者の音声への変換を行い，その変換音声に対してスコアリングを行う．ES-to-Speech によって得られる変換音声は，話速やイントネーションは事前収録出力話者に関わらず同一であり，制御可能な声質のみが個々の話者に依存する．そのため，変換音声に対してスコアリングすることで，より声質制御に適したスコアが得られるものと考えられる．

4.3 カーネル回帰 GMM （Kernel Regression GMM: KR-GMM）に基づく ES-to-Speech 変換音声の声質制御法

従来の MR-GMM では，重回帰分析により，声質表現語スコアベクトルから各 GMM の出力平均ベクトルへの線形変換を行うことで，スコア操作による声質の制御を実現している．そのため，声質表現語スコアベクトルと出力平均ベクトルが非線形な関係性を持つ場合，声質の制御性能が劣化すると考えられる．そこで，本稿では，カーネル回帰による非線形な回帰モデルの導入を行う．カーネル回帰とは，説明変数を高次元空間へ写像し，回帰を行う手法である．モデルの表現能力を抑える正則化を行うことで，過学習を防ぐことも可能である．

カーネル回帰では，重み $\boldsymbol{w}$ が与えられた際の m 番目の分布に対する出力平均ベクトル $\boldsymbol{\mu}_m^{(Y)}(\boldsymbol{w})$ の d 次元目の要素 $\mu_{m,d}^{(Y)}(\boldsymbol{w})$ を

$$\mu_{m,d}^{(Y)}(\boldsymbol{w}) = \sum_{j=1}^{J} \boldsymbol{v}_{j,m,d}^{\top} \phi(w_j) \tag{3}$$

で表す．ここで，$\phi(\cdot)$ は高次元空間へ写像する関数を表す．また，$\boldsymbol{v}_{j,m,d}$ は高次元空間におけるベクトルであり，学習データとして用いる各事前収録出力話者に対する声質表現語スコアベクトル $\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(S)}$ を用いて以下で表される．

$$\boldsymbol{v}_{j,m,d} = \sum_{s=1}^{S} \alpha_{j,m,d,s} \phi(w_j^{(s)}) \tag{4}$$

ここで，S は事前収録出力話者数である．式 (4) を式 (3) に代入すると，

$$\mu_{m,d}^{(Y)}(\boldsymbol{w}) = \sum_{j=1}^{J} \boldsymbol{\alpha}_{j,m,d} \boldsymbol{k}_j(w_j) \tag{5}$$

$$\boldsymbol{k}_j(w_j) = [k(w_j^{(1)}, w_j), \cdots, k(w_j^{(S)}, w_j)]^{\top} \tag{6}$$

が得られる．ここで，$\boldsymbol{\alpha}_{j,m,d} = [\alpha_{j,m,d,1}, ..., \alpha_{j,m,d,S}]$ であり，$k(\cdot,\cdot)$ はカーネル関数を表す．さらに，$\boldsymbol{\alpha}_{m,d} = [\boldsymbol{\alpha}_{1,m,d}, \cdots, \boldsymbol{\alpha}_{J,m,d}]$ および $\boldsymbol{k}(\boldsymbol{w}) = [\boldsymbol{k}_1^{\top}(w_1), \cdots, \boldsymbol{k}_J^{\top}(w_J)]^{\top}$ とおくと，式 (5) は次式で表される．

$$\mu_{m,d}^{(Y)}(\boldsymbol{w}) = \boldsymbol{\alpha}_{m,d} \boldsymbol{k}(\boldsymbol{w}) \tag{7}$$

なお，本稿では，カーネル関数として，次式で示すガウスカーネルを用いる．

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \exp(-\beta ||\boldsymbol{x} - \boldsymbol{x'}||) \tag{8}$$

ここで，β は任意の値をとるパラメータである．

各事前収録出力話者に対する出力平均ベクトル $\boldsymbol{\mu}_m^{(Y,1)}, \cdots, \boldsymbol{\mu}_m^{(Y,S)}$ と声質表現語スコアベクトル $\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(S)}$ を用いて，$\boldsymbol{\alpha}_{m,d}$ を最適化する．全事前収録出力話者に対する出力平均ベクトルの d 次元目の要素からなるベクトルを $\boldsymbol{\mu}_{m,d}^{(1:S)} = [\mu_{m,d}^{(Y,1)}, \cdots, \mu_{m,d}^{(Y,S)}]$ とすると，カーネル回帰における自乗誤差に基づく評価関数は次式で表される．

$$\epsilon^2 = (\boldsymbol{\mu}_{m,d}^{(1:S)} - \boldsymbol{\alpha}_{m,d} \boldsymbol{K}^{(1:S)})(\boldsymbol{\mu}_{m,d}^{(1:S)} - \boldsymbol{\alpha}_{m,d} \boldsymbol{K}^{(1:S)})^{\top} + \lambda \boldsymbol{\alpha}_{m,d} \boldsymbol{\alpha}_{m,d}^{\top} \tag{9}$$

ここで，$\boldsymbol{K}^{(1:S)} = [\boldsymbol{k}(\boldsymbol{w}^1), \cdots, \boldsymbol{k}(\boldsymbol{w}^S)]$ である．また，右辺第 2 項は正則化項であり，正則化パラメータ λ により，モデルの表現能力の調整を行う．式 (9) を最小化する $\hat{\boldsymbol{\alpha}}_{m,d}$ は次式で表される．

$$\hat{\boldsymbol{\alpha}}_{m,d} = \boldsymbol{\mu}_{m,d}^{(1:S)} \boldsymbol{K}^{(1:S)^{\top}} (\boldsymbol{K}^{(1:S)} \boldsymbol{K}^{(1:S)^{\top}} + \lambda \boldsymbol{I})^{-1} \tag{10}$$

式 (10) を式 (7) に代入することで，声質表現語スコアベクトル $\boldsymbol{w}$ が与えられた際の平均ベクトルの各要素は次式で得られる．

$$\mu_{m,d}^{(Y)}(\boldsymbol{w}) = \boldsymbol{\mu}_{m,d}^{(1:S)} \boldsymbol{K}^{(1:S)^{\top}} (\boldsymbol{K}^{(1:S)} \boldsymbol{K}^{(1:S)^{\top}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{k}(\boldsymbol{w}) \tag{11}$$

本稿では，カーネル関数のパラメータ β および正則化パラメータ λ は Cross Validation により最適化する．

5. 実験的評価

5.1 実 験 条 件

MR-GMM の学習のために，元話者として男性喉頭摘出者 1 名の食道音声を，出力話者として女性 27 名，男性 34 名（計 61 名）の健常者音声を用いる．発話内容は全話者とも同一の音素バランス文 40 文である．スコアリング対象音声は，61 名の健常者の自然音声 (NS) と，食道音声から 61 名の健常者音声へそれぞれ ES-to-Speech を行った変換音声 (ES2NS) である．スコアリングは男性 1 名が行う．声質の評価項目は声質表現語対[8] の内の以下の 7 つであり，-3 から 3 までの 7 段階評定 (例; 非常に男性的な声: -3，どちらでもない: 0，非常に女性的な声: 3) でスコアリングを行う．

- 男性的な声－女性的な声 (gender)
- かすれた声－澄んだ声 (clearness)
- 老けた声－若い声 (age)
- 太い声－細い声 (deepness)
- 張りのない声－張りのある声 (forcefulness)
- 迫力のある声－弱々しい声 (powerfulness)
- 落ち着きのある声－落ち着きのない声 (calmness)

各声質表現語対ごとに平均 0，分散 1 に標準化されたスコアを，声質表現語スコアベクトルの要素とする．

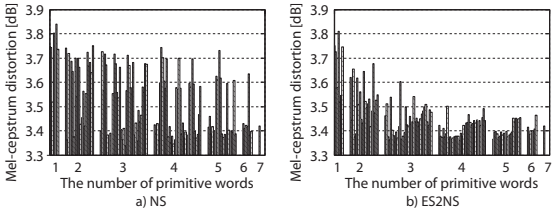

図 1 スペクトルの予測誤差
Fig. 1 Prediction error of spectrum.

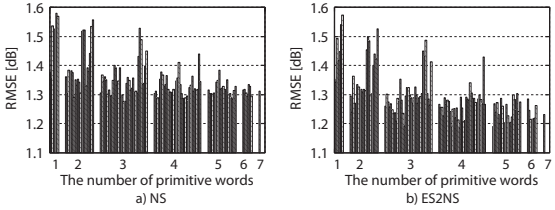

図 2 非周期成分の予測誤差
Fig. 2 Prediction error of aperiodic components.

スペクトル特徴量として，0 次から 24 次のメルケプストラム係数を用いる．音源特徴量としては STRAIGHT[10] によって抽出された対数 F_0 と 5 帯域 (0-1, 1-2, 2-4, 4-6, 6-8 kHz) で平均化された非周期成分を用いる．スペクトル及び非周期成分推定用 MR-GMM の混合数は 64 とする．F_0 推定用の特定話者 GMM の混合数は 32 とする．

客観評価尺度として，回帰分析における予測誤差を用いる．予測誤差の計算時には，学習話者を 60 名，評価話者を 1 名とする Cross Validation を行う．スペクトル推定用 GMM の出力平均ベクトルと非周期成分推定用 GMM の出力平均ベクトルに対する誤差尺度として，それぞれメルケプストラムひずみ（Mel-CD）と，二乗平均平方根誤差（RMSE）を用いる．また，対数 F_0 の平均に対する誤差尺度として，RMSE を用いる．

5.2 スコアリング対象音声についての検討

自然音声に対するスコアと，ES-to-Speech 変換音声に対するスコアを比較し，どちらが声質制御に適しているかを調査する．

自然音声に対するスコア（NS）と，ES-to-Speech 変換音声（ES2NS）に対するスコアを用いて重回帰分析を行い，それぞれの予測誤差を比較する．なお，声質表現語対を 1 つのみ用いた場合，7 つ全て用いた場合など，全組み合わせ（$2^7 - 1 = 127$ 通り）で予測誤差を求める．客観評価結果を**図 1**，**図 2**，**図 3** に示す．

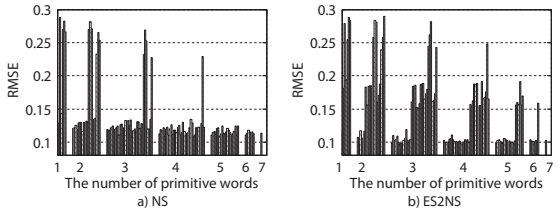

図 3 対数 F_0 の予測誤差
Fig. 3 Prediction error of logarithmic F_0.

予測誤差が小さい組み合わせに注目すると，ES-to-Speech 変換音声に対するスコアを用いることで，特に非周期成分，対数 F_0 において予測誤差が減少することが分かる．このことから，変換音声に対するスコアリングの有効性が確認できる．なお，スペクトルにおいては，予測誤差が最小となる声質表現語対セットを用いた際には，自然音声に対するスコアと ES-to-Speech 変換音声に対するスコア間で大きな差は見られないが，ES-to-Speech 変換音声を用いることで，声質表現語対の数を増やすことで予測誤差が小さくなる傾向がより明確となる．

また，gender，clearness，age，deepness，forcefulness の組み合わせを用いた場合，全ての特徴量において，低い予測誤差（下位 10% ）を示している．よって，以降の実験では，これら 5 つの声質表現語対を用いる．

5.3 KR-GMM に基づく声質制御法

MR-GMM と KR-GMM を比較し，どちらが声質制御に適しているかを調査する．学習には，ES-to-Speech 変換音声に対するスコアを用いる．

5.3.1 客観評価

カーネル回帰分析において，パラメータ λ，β を変化させた際の予測誤差を**図 4**，**図 5**，**図 6** に示す．比較のため，重回帰分析による結果も示す．カーネル回帰分析では，パラメータ λ 及び β の設定が大きく予測誤差に影響を与えることが分かる．図 4，図 6 より，スペクトルと対数 F_0 に関しては，カーネル回帰分析において適切にパラメータを調整することで，重回帰分析と比較してより良い予測精度が得られることが分かる．一方で，図 5 より，非周期成分に関しては，カーネル回帰分析により重回帰分析を上回る予測精度は得られないことが分かる．

5.3.2 主観評価

MR-GMM と KR-GMM の声質制御性能を評価するため，変換音声の主観評価実験を行う．カーネル回帰分析におけるパラメータ λ，β は，各特徴量において 5.3.1 で求めた最適値に設定する．

声質制御の操作性能を評価するために，声質表現語スコアベクトルの中で，deepness と forcefulness

Vol.2011-SLP-85 No.11
2011/2/5

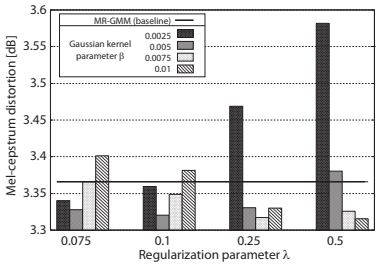

図 4 カーネル回帰におけるスペクトルの予測誤差
Fig. 4 Prediction error of spectrum.

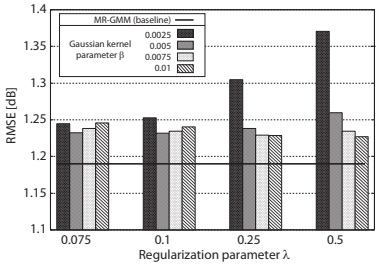

図 5 カーネル回帰における非周期成分の予測誤差
Fig. 5 Prediction error of aperiodic components.

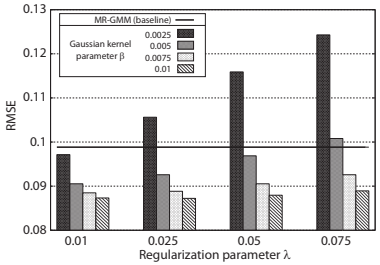

図 6 カーネル回帰における対数 F_0 の予測誤差
Fig. 6 Prediction error of logarithmic F_0.

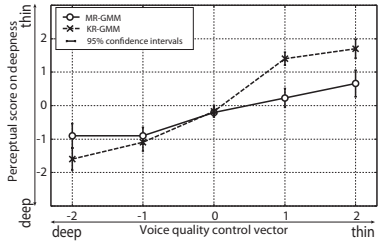

図 7 deepness に関する主観評価結果
Fig. 7 Perceptual score on deepness.

に関するスコアを操作して，変換音声の声質制御を行う．被験者は，参照音声と刺激音声を比較し，各声質表現語対に対するスコア（-2〜2 の 5 段階）を用いて，刺激音声の声質を評価する．参照音声として，声質表現語スコアベクトルの値をすべて 0 に設定した際の変換音声を用いる．刺激音声として，対象とする声質表現語スコアベクトルの値を-2〜2 の 5 段階で変化させた場合の変換音声を用いる．その際に，他の声質表現語スコアベクトルの値は 0 に固定する．また，声質制御時における変換音声の自然性を評価するために，gender に関して声質制御を行った際の変換音声の自然性を，1 から 5 の 5 段階オピニオンスコアによるオピニオンテストにより評価する．両評価において，評価文は，学習データに含まれない 5 文であり，被験者は MR-GMM 学習用のスコアリングを行った者を含まない男女 10 名である．

図 7 及び**図 8** に操作性能に関する主観評価結果を示す．図 8 において，MR-GMM 及び KR-GMM により声質制御された変換音声に対するスコアは，制御時に設定したスコアと強い正の相関があることから，声質制御が適切に行われていることが分かる．また，特に図 7 において，KR-GMM による変換音声に対するスコアは，MR-GMM のものと比較し，より広い範囲をカバーしており，より設定したスコアに近いものとなっている．このことから KR-GMM は，より適切に声質制御を行えることが分かる．

次に，自然性に関する主観評価結果を**図 9** に示す．声質表現語スコアベクトルの値を極端に大きくしたり小さくしたりすると，自然性が劣化する傾向が見られる．これは，学習データがほとんど存在しないスコア領域であるためである．また，声質表現語スコアベクトルの値が-3 から-5 のときに注目すると，KR-GMM の方が MR-GMM よりも自然性の劣化が小さい．このことから，カーネル回帰を用いて声質表現語スコアと音声モデルパラメータの非線形な関係をモデル化することで，より頑健な声質制御が可能であることが分かる．

これらの結果から，KR-GMM を用いることで，より高い声質制御性能が得られることが分かる．

Vol.2011-SLP-85 No.11
2011/2/5

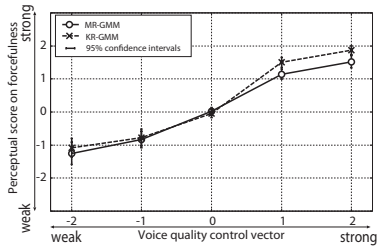

図 8 forcefulness に関する主観評価結果
Fig. 8 Perceptual score on forcefulness.

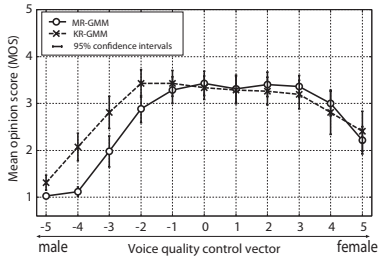

図 9 自然性に関する主観評価結果
Fig. 9 Mean opinion score on naturalness.

6. ま と め

統計的声質変換に基づく食道音声強調（Esophageal Speech-to-Speech: ES-to-Speech）において，直感的な声質制御を実現するために，重回帰混合正規分布モデル（multiple regression Gaussian mixture model: MR-GMM）に基づく ES-to-Speech を提案した．また，声質制御性能を改善するために，ES-to-Speech 変換音声を用いたスコアリング手法とカーネル回帰 GMM（Kernel Regression GMM: KR-GMM）に基づく声質制御法を提案した．実験結果から，提案法により，ES-to-Speech において直感的な声質制御が可能であることを示した．また，ES-to-Speech 変換音声を用いたスコアリング手法の有効性を客観的に示すとともに，KR-GMM は MR-GMM よりも高い声質制御性能が得られることを示した．

謝辞 本研究の一部は，科研費補助金若手研究（A）及び総務省 SCOPE により実施したものである．

参 考 文 献

1) Y. Stylianou, O. Cappe, and E. Moulines.: Continuous probabilistic transform for voice conversion, *IEEE Trans. Speech and Audio Processing*, Vol.6, No.2, pp.131–142 (1998).
2) 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 (2007).
3) H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano.: Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models, *IEICE Transactions on Information and Systems*, Vol.E93-D, No.9, pp.2472–2482 (2010).
4) T. Toda, Y. Ohtani, and K. Shikano.: One-to-many and many-to-one voice conversion based on eigenvoices, *Proc. ICASSP*, pp.1249–1252 (2007).
5) K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano.: Regression Approaches to Voice Quality Control Based on One-to-Many Eigenvoice Conversion, *Proc. 6th ISCA Speech Synthesis Workshop*, pp.101–106 (2007).
6) A. Kain and M.W. Macon.: Spectral voice conversion for text-to-speech synthesis, *Proc. ICASSP*, pp.285–288 (1998).
7) T. Toda, A.W Black, K. Tokuda.: Statistical Mapping between Articulatory Movements and Acoustic Spectrum with a Gaussian Mixture Model , *Speech Communication*, Vol.50, No.3, pp.215–227 (2008).
8) 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, 音響誌, Vol.55, No.6, pp.405–411 (1999).
9) 木戸博, 粕谷英樹.: 通常発話の声質に関連した日常表現語 聴取評価による抽出, 音響誌, Vol.57, No.5, pp.337–344 (2001).
10) H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, *Speech Communication*, Vol.27, No.3-4, pp.187–207 (1999).