中間話者コーパスを用いたアニメーション演技音声のための話者変換
Voice conversion for animation acting voice with intermediate speaker corpus

塩出萌子
法政大学情報科学部ディジタルメディア学科
E-mail: moeko.shiode.i4@stu.hosei.ac.jp

概要

This paper proposes a Gaussian mixture model (GMM)-based voice conversion (VC) method for animation acting voice using intermediate speaker corpus. Although mapping parameter estimation requires a lot of parallel data, in VC between voice actors, it is difficult to collect parallel data. In order to resolve the problem, two mapping functions, which trained with intermediate speaker corpuses, are used. The corpuses are consisted of two parallel corpuses; source (new voice actor) -to-intermediate and intermediate-to-target (old voice actor). By the intermediate speaker speaks sentences, which spoken in animations by each actor, the two parallel corpuses are created. Three opinion test for speaker individuality, intelligibility and speech quality were conducted, and these test showed speaker individuality was 50% from ABX experiments, intelligibility was 4.1 point from mean opinion score (MOS) experiments and speech quality was 3.7 point from MOS experiments.

1 まえがき

長期放映されているアニメーションは，大人から子供まで幅広い世代で認知度が高く，楽しまれている．そのような作品に対し，視聴者は，各登場人物役の外見や性格、声色などについて固定されたイメージを持つ．そのため、年代の経過によって声優が交代した際，声質や表現方法などの差異が，視聴者の今まで持っていた印象に違和感を与える．これらの違和感は，視聴者の番組離れの原因となるため，問題となる．

そこで本論文では，中間話者パラレルコーパスを用いて，新声優の声を旧声優の声に変換するための話者変換器の作成を目指す．元話者 (新声優) の特徴量を，目標話者 (旧声優) の対応する特徴量と変換し，変換した音声が目標話者の音声と聴覚的に違和感がないことを目標とする．

2 話者変換

音声とは，腹筋が横隔膜を押し上げることによって肺から押し出された空気が声帯を周期的に震わせて音を出し，次に，口の形や舌の位置で生成したい音韻に合わせて音色を決め，最後に口から空気中に放射させることで生成される．

音声により表現・伝達される情報は、言語的情報 (言語情報)，パラ言語的情報，非言語的情報 (個人性情報) がある．言語的情報とは，文字言語に直接含まれるものを指す．パラ言語的情報とは，直接的に文字表現に含まれないが，韻律的特徴によって話者の発話意図や感情状態などが表現されるもので，話者がその表現を制御できるものを指す．非言語的情報とは，発話内容に関係せず，話者が意識的に制御し得ない個人的特徴や身体状態などの情報を指す．

音声信号を分析し，これらの情報を表すそれぞれの特徴パラメータを取り出し，それに基づいて処理を行うことを一般に音声情報処理と呼ぶ．音声波形から言語情報を取り出すことを音声認識，個人性情報を取り出すことを話者認識と呼ぶ．ここでは，元話者と目標話者の個人性情報を取り出し変換を行う，話者変換を扱う．

2.1 音声の個人性情報

話者変換は，音声の個人性情報を変換する．音声の個人性は，音源特性，声道特性，韻律特性からなる．

2.1.1 音源特性

音源特性は，声帯振動を表したものである．声帯の緊張が大きく，かつ肺からの空気圧が高いと，声帯の開閉周期，すなわち振動周期が短く，音源の高さが高くなり，逆の場合は低くなる．これが音声の高さ (ピッチ) に対応する．声帯の振動周期のことを基本周期，その逆数を基本周波数 (F_0) と呼ぶ．

2.1.2 声道特性

声道特性は，口の形や舌の位置など発声器官の形状で決まる音色をスペクトル包絡で表したものである。波形の形，つまり音色を変化させる声道フィルターの役割がある．

2.1.3 韻律特性

韻律特性は，音節，拍，アクセント，イントネーション，リズム，速さ，時間長，ポーズ (休止)，強弱などがあげられる．

2.2 声質変換

本研究の変換対象は，同じキャラクターを演じている新旧の声優である．新声優は，旧声優の演技 (F_0 の変化や話速などの韻律的特徴) を模倣しているとは限らない．しかし演じている登場人物は同一のため，各声優のキャラクターの性格・感情表現の方法，すなわちキャラクター性は同一とみなせる．つまり，パラ言語情報から伝達されるキャラクター性は同一と考えられる．またアニメーション作品では，肉体的な要因である声質などの非言語情報が視聴者のイメージと異なることにより，違和感が生じると言われている [9]．そこで本論文では，特徴量として声道特性に着目したスペクトル変換を行う．

声質変換とは，入力と出力の対応関係を記述する変換モデルに基づき，言語情報を保ちながら話者性などの非言語情報を変換する技術である．統計的手法に基づく従来研究として，混合正規分布モデル (GMM) に基づく変換法がある [2]．GMM で 2 話者間の特徴量 $\boldsymbol{x}$ と $\boldsymbol{y}$ の対応関係をモデル化し，元話者の特徴量に対応する目標話者の特徴量を推定する．

2.2.1 GMM を用いたスペクトル変換

■学習 フレーム t において，入力特徴量 (元話者) を $\boldsymbol{x}_t$、出力特徴量 (目標話者) を $\boldsymbol{y}_t$ とする．$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ は D 次元のベクトルである．GMM のモデルパラメータは，結合ベクトル $\boldsymbol{z}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}$ を用いた式 (1) の結合確率密度をモデル化し，推定する．

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{n=1}^{N} w_n \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_n^{(z)}, \boldsymbol{\Sigma}_n^{(z)}) \tag{1}$$

$$\boldsymbol{\mu}_n^{(z)} = \begin{bmatrix} \mu_n^{(x)} \\ \mu_n^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_n^{(z)} = \begin{bmatrix} \Sigma_n^{(xx)} & \Sigma_n^{(xy)} \\ \Sigma_n^{(yx)} & \Sigma_n^{(yy)} \end{bmatrix} \tag{2}$$

$\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトル μ と共分散行列 Σ で表される正規分布である．n は分布番号，N は混合数，w_n は n 番目の分布の重みを表す．

Supervisor: Prof. Katunobu Itou

■フレームベース変換　$\boldsymbol{x}_t$ が与えられた場合の $\boldsymbol{y}_t$ の条件付き確率密度を GMM でモデル化する (式 (3)).

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(Z)}) = \sum_{n=1}^{N} P(n|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)})P(n|\boldsymbol{y}_t,\boldsymbol{\lambda}^{(z)}) \quad (3)$$

ここで，$P(n|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(Z)})$ は重み，$P(\boldsymbol{y}_t|\boldsymbol{x}_t,n,\boldsymbol{\lambda}^{(Z)})$ は正規分布で，式 (4)(5) のように表せる．

$$P(n|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(Z)}) = \frac{w_n\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_n^{(x)},\boldsymbol{\Sigma}_n^{(xx)})}{\sum_{n=1}^{N} w_n\mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_n^{(x)},\boldsymbol{\Sigma}_n^{(xx)})} \quad (4)$$

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t,n,\boldsymbol{\lambda}^{(Z)}) = \mathcal{N}(\boldsymbol{y}_t;\boldsymbol{E}_{\boldsymbol{n},\boldsymbol{t}}^{(\boldsymbol{y})},\boldsymbol{D}_n^{(y)}) \quad (5)$$

フレーム t における n 番目の分布の条件付き確率密度の平均ベクトル $\boldsymbol{E}_{n,t}^{(y)}$，共分散行列 $\boldsymbol{D}_n^{(y)}$ は式 (6)(7) で表せる．

$$\boldsymbol{E}_{n,t}^{(y)} = \boldsymbol{\mu}_n^{(y)} + \boldsymbol{\Sigma}_n^{(yx)}\boldsymbol{\Sigma}_n^{(xx)^{-1}}(\boldsymbol{x}_t - \boldsymbol{\mu}_n^{(x)}) \quad (6)$$

$$\boldsymbol{D}_n^{(y)} = \boldsymbol{\Sigma}_n^{(yy)} - \boldsymbol{\Sigma}_n^{(yx)}\boldsymbol{\Sigma}_n^{(xx)^{-1}}\boldsymbol{\Sigma}_n^{(xy)} \quad (7)$$

最小平均二乗誤差に基づく変換では，$\boldsymbol{x}_t$ から推定される変換特徴量 $\hat{\boldsymbol{y}}_t$ を式 (8) で表せる．

$$\hat{\boldsymbol{y}}_t = \sum_{n=1}^{N} P(n|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)})\boldsymbol{E}_{n,t}^{(y)} \quad (8)$$

この変換処理は，フレームごとに独立に行われる．このため，時間方向の相関を考慮していない特徴量遷移が発生する．また，統計的な処理により音声信号が持つ詳細な特徴が過剰に平滑化されてしまう．スペクトル構造の詳細が消失は，変換音声の品質劣化を引き起こす．そこで，動的特徴量を導入し，パラメータの系列内変動を考慮した最尤変換を行うことで，上記の問題解決を実現する [3]．最尤変換とは，条件付確率密度分布の尤度最大化に基づいて変換特徴量 $\hat{\boldsymbol{y}}_t$ を推定する．

$$\hat{\boldsymbol{y}}_t = \mathrm{argmax} P(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\lambda}) \quad (9)$$

■動的特徴量を考慮した最尤変換　D 次元の静的特徴量と動的特徴量からなる $2D$ 次元の入力・出力特徴量ベクトルを $\boldsymbol{X}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \Delta\boldsymbol{x}_t^{\mathrm{T}}]$，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \Delta\boldsymbol{y}_t^{\mathrm{T}}]$ とする．最尤変換は，静的・動的特徴量間の明示的な関係を考慮し，条件付確率密度関数 $P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)})$ の尤度最大化に基づき，変換特徴量 $\hat{\boldsymbol{y}}_t$ を求める．静的・動的特徴量系列 $\boldsymbol{Y}_t$ は，静的特徴量系列 $\boldsymbol{y}_t$ の線形変換として式 (10) と表現できる．(図 1)

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (10)$$

ここで，$\boldsymbol{W}$ は静的特徴量系列を静的・動的特徴量系列に拡張する変換行列である．

単一分布系列 $\boldsymbol{n} = [n_1, n_2, ..., n_t, ..., n_T]^{\mathrm{T}}$ による近似で尤度関数を最大とする $\hat{\boldsymbol{y}}$ 推定する．式 (11) で最尤分布系列 $\hat{\boldsymbol{n}}$ を決定し，変換特徴量系列 $\hat{\boldsymbol{y}}$ を決定する．

$$\hat{\boldsymbol{n}} = \mathrm{argmax} P(\boldsymbol{n}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}) \quad (11)$$

$$\begin{aligned}\hat{\boldsymbol{y}} &= \mathrm{argmax} P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}) \\ &\simeq \mathrm{argmax} P(\hat{\boldsymbol{n}}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)})P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{n}},\boldsymbol{\lambda}^{(Z)}) \\ &= (\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}_{\hat{\boldsymbol{n}}}^{(Y)^{-1}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}_{\hat{\boldsymbol{n}}}^{(Y)^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{n}}}^{(Y)}\end{aligned} \quad (12)$$

ここで，$\boldsymbol{E}_{\hat{\boldsymbol{n}}}^{(Y)}$，$\boldsymbol{D}_{\hat{\boldsymbol{n}}}^{(Y)^{-1}}$ は，式 (13)(14) である．

$$\boldsymbol{E}_{\hat{\boldsymbol{n}}}^{(Y)} = [\boldsymbol{E}_{\hat{\boldsymbol{n}}_{1,1}}^{(Y)}, \boldsymbol{E}_{\hat{\boldsymbol{n}}_{2,2}}^{(Y)}, ..., \boldsymbol{E}_{\hat{\boldsymbol{n}}_{t,t}}^{(Y)}, ..., \boldsymbol{E}_{\hat{\boldsymbol{n}}_{T,T}}^{(Y)}] \quad (13)$$

$$\boldsymbol{D}_{\hat{\boldsymbol{n}}}^{(Y)^{-1}} = \mathrm{diag}[\boldsymbol{D}_{\hat{\boldsymbol{n}}_{1,1}}^{(Y)^{-1}}, \boldsymbol{D}_{\hat{\boldsymbol{n}}_{2,2}}^{(Y)^{-1}}, ..., \boldsymbol{D}_{\hat{\boldsymbol{n}}_{t,t}}^{(Y)^{-1}}, ..., \boldsymbol{D}_{\hat{\boldsymbol{n}}_{T,T}}^{(Y)^{-1}}] \quad (14)$$

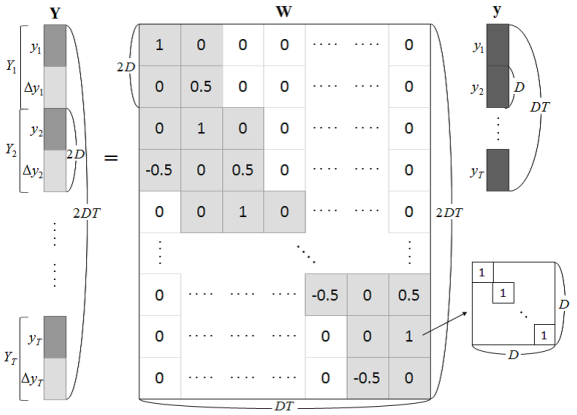

図 1. 静的特徴量と動的特徴量の関係

■系列内変動を考慮した最尤変換　出力特徴量の系列内変動 (Global Variance:GV) を式 (15) で表す．

$$\boldsymbol{v}(\boldsymbol{y}) = [v(1), v(2), ..., v(d), ..., v(D)]^{\mathrm{T}} \quad (15)$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}(y_t(d) - \frac{1}{T}\sum_{t=1}^{T}y_t(d))^2 \quad (16)$$

$v(d)$ は d 次元目の GV である．系列内に含まれる全フレームにわたって計算される GV を用いて GV の確率密度 $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ をモデル化する．GV を考慮した最尤変換は，上述の最尤変換に GV に関する制約条件の下での尤度最大化を行う．変換時には式 (17) の対数尤度関数を最大化する．

$$\mathcal{L} = \log P(\hat{\boldsymbol{n}}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)})P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{n}},\boldsymbol{\lambda}^{(Z)})^{\omega}P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) \quad (17)$$

$\mathcal{L}$ を最大化する $\hat{\boldsymbol{y}}$ を求めるために，式 (18) の一次導関数を計算する．

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{y}} &= \omega(-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}_{\hat{\boldsymbol{n}}}^{(Y)^{-1}}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}_{\hat{\boldsymbol{n}}}^{(Y)^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{n}}}^{(Y)}) \\ &\quad + [\boldsymbol{v}_1'^{\mathrm{T}}, \boldsymbol{v}_2'^{\mathrm{T}}, ..., \boldsymbol{v}_t'^{\mathrm{T}}, .., \boldsymbol{v}_T'^{\mathrm{T}}]^{\mathrm{T}}\end{aligned} \quad (18)$$

$$\boldsymbol{v}_t' = [\boldsymbol{v}_t'(1), \boldsymbol{v}_t'(2), ..., \boldsymbol{v}_t'(d), ..., \boldsymbol{v}_t'(D)]^{\mathrm{T}} \quad (19)$$

$$\boldsymbol{v}_t'(d) = -\frac{2}{T}\boldsymbol{p}_v^{(d)^{\mathrm{T}}}(\boldsymbol{v}(\hat{\boldsymbol{y}}) - \boldsymbol{\mu}_v)(\hat{y}_t(d) - \bar{\tilde{y}}(d)) \quad (20)$$

ここで，$\boldsymbol{p}_v^{(d)}$ は共分散行列 Σ_v の逆行列 P_v における d 番目のベクトルを表す．$\hat{\boldsymbol{y}}$ は，式 (21) を用いて繰り返し更新し，求める．

$$\hat{\boldsymbol{y}}^{(i+1)-th} = \hat{\boldsymbol{y}}^{(i)-th} + \alpha\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{y}}\right|_{\boldsymbol{y}=\boldsymbol{y}^{(i)-th}} \quad (21)$$

2.2.2　学習データ

変換モデルの学習の際には，学習データとしてパラレルコーパスを必要とする．パラレルコーパスとは，同じ発話内容の文で構成される同一発話文セットである．これは，元話者と目標話者が同じ発話内容であることで，対応が取りやすく，時間変化に対しても安定した学習が期待でき，学習精度が上がるためである．前述の声質変換は，元話者と目標話者のパラレルコーパスが存在する場合に構築可能な話者変換である．

3　中間話者コーパスを用いた変換方法

パラレルコーパスの構築には，元話者と目標話者の発話内容を制御した音声収録をする必要がある．しかし本論文で扱う話者変換の変換対象は声優のため，発話内容を制御したデータ収集が困難で，必要なデータを自由に作成できず，必ずしも元話者と目標話者のパラレルコーパスを用意できない．

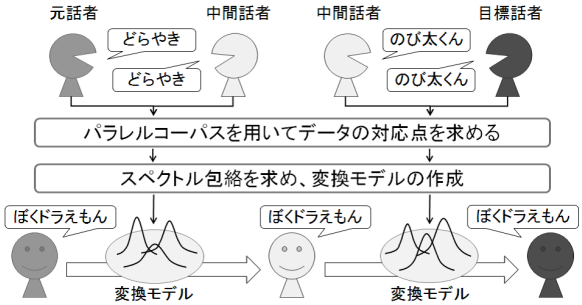

図 2. システム概要

そこで，発話内容を自由に制御可能な中間話者 $\boldsymbol{m}$ を考え，元話者と目標話者の特徴量が独立と仮定すると，目標話者と中間話者の尤度最大化は以下のようになる．

$$\underset{\boldsymbol{y},\boldsymbol{m}}{\operatorname{argmax}}\, p(\boldsymbol{y},\boldsymbol{m}|\boldsymbol{x}) \approx \underset{\boldsymbol{y},\boldsymbol{m}}{\operatorname{argmax}}\, p(\boldsymbol{y}|\boldsymbol{m})p(\boldsymbol{m}|\boldsymbol{x}) \tag{22}$$

すなわち，元話者から中間話者への声質変換と，中間話者から目標話者への声質変換を多段的に行うことで，元話者と目標話者の変換モデルを用いずに声質変換ができることを意味する．

提案法の概要を図 2 に示す．まず，元話者と中間話者，中間話者と目標話者でそれぞれパラレルコーパスを作成する．そして，作成したパラレルコーパスを用いて，それぞれの変換モデルを学習する．変換手順は，まず，元話者の音声データを入力とし，元話者と中間話者で学習した変換モデルを用いて中間話者の音声へ変換する．そして，変換した中間話者音声データから，中間話者と目標話者で学習した変換モデルを用いて，目標話者の音声へと変換し出力する．変換の際は，まず，入力データ (元話者データ) から声道特性を推定し，変換モデルを用いて中間話者の声道特性へと変換する．そして，変換した中間話者の声道特性から変換モデルを用いて目標話者の声道特性へと変換する．声道特性以外の成分 (声の高さ，話速，イントネーションなど) は残差である．今回は演技の方法は変換しなくて良いとしているため，元話者データの残差に，求めた目標話者の声道特性を畳み込み，再合成し出力する．

3.1 データ収集

アニメーション中の BGM のない話者単独の音声部分からデータ収集を行う．本稿では，長期放映で代表的なアニメ「ドラえもん」から音声データを収集する．ドラえもんは，1979 年 4 月から現在まで放映されているが，主要登場人物の声優が 2005 年 4 月に交代した．表 1 は，収集に用いた音声データのエピソード数と全データ時間，表 2 は，ドラえもんの旧声優と新声優の一覧である．

表 1. 収集データ

	旧声優	新声優
時間 (エピソード)	613 分 42 秒 (47)	513 分 17 秒 (50)

表 2. 声優一覧

キャラクター	旧声優	現声優
①ドラえもん	大山のぶ代	水田わさび
②野比のび太	小原乃梨子	大原めぐみ
③源静香	野村道子	かかずゆみ
④骨川スネ夫	肝付兼太	関智一
⑤剛田武	たてかべ和也	木村昴

3.2 パラレルコーパスの作成

収集した新旧の声優の音声データに対応する中間話者データを作成し，元話者，目標話者それぞれについてパラレルコーパスを作成する．声優のアニメーション中の発声は，話速は速く，声による感情表現が豊かで，一語一語はっきりと発音されている特徴があった．そこで，良い変換モデルの作成のために，発話内容，発話口調 (アクセント・イントネーションなど)，発話時間 (話速)，F_0 の変化に注意し，対応のとりやすい中間話者データの作成を心掛け，録音を行った．また，中間話者が少しでも模倣しやすくすることをねらい，驚きや怒り，悲しみ (泣き声) のセリフは除き，元話者と目標話者のデータ条件は，感情表現の激しくない発声の音声データとした．

3.3 音素バランスを考慮したパラレルコーパス

中間話者パラレルコーパスを学習データとし，変換モデルを作成する．従来の声質変換では，音声データとして ATR 日本語音声データベース [5] が用いられる．これは，新聞，雑誌，小説，教科書等の文献から無作為に抽出した約一万の文をもとに，音素環境をバランスされて作成した朗読音声データベースである．

しかし，今回の変換対象は演技音声である．朗読音声がテキストの伝達を目的とした文章の文法的な構造であるのに対し，演技音声はテキスト (作品) の表現を目的とした人物と人物の対話構造である．また，会話文法には比較的に短い文章が多く，同じ語を繰り返さないため省略することや，文末に “さ”，“ね”，“よ” 等の終助詞がつくことが多く，また語尾を伸ばすこと，“えー” や “あー” などの言い淀み，頻出率が高い単語 (登場人物の名前など)，言い誤りや言い直しなどがある．音響モデル構築に用いるデータは，日本語の発声に含まれうる音声現象を可能な限り多く含んだ学習データが望ましい [4]．

そこで，ATR 音声データベースの各音節の頻度を参考にする．変換対象は声優であり，必要となるデータを自由に収集出来ない．そのため，まずは収集した全データの音節を調べる．そして調査の結果，ATR 音声データベースの頻度より下回る音節を含むデータは全て学習データに含むことで，音韻バランスを考慮したパラレルコーパスを作成する．

3.4 DP マッチングによる対応点決定

学習データから変換モデルを作成する際に，パラレルデータの対応点を求める．これには，動的時間伸縮法 (Dynamic Time Warping:DTW) を用いた [6]．ケプストラム距離を距離コストとして変換する二者間の距離を評価し，最短経路を選択することで対応点を決定する．

4 評価実験

提案手法の変換精度を評価した．評価対象はドラえもんの 3 キャラクター (ドラえもん，のび太，スネ夫) とした．変換モデルは，新声優 (元話者) と中間話者のパラレルコーパス，中間話者と旧声優 (目標話者) のパラレルコーパス共に 50 文で作成した．各キャラクターの 50 文に含まれる音節数は，表 3 に示す．特徴量は，20 次元の線形予測係数を線スペクトル対に変換したものとし，GMM は 64 混合で構築した．フレームシフトは 5ms，分析窓長は 512ms，音声のサンプリング周波数は 16kHz にダウンサンプリングした．被験者は 10 名であり，音声はヘッドホンにより提示し，聞き返しはなしとした．評価に用いる音声データは，学習データに含まれない発話時間が 1 秒以上のものとし，被験者ごとにランダムに提示した．また，喜怒哀楽による偏りをなくすため，怒りや悲しみ以外の感情表現の激しくない発声の音声データとした．

表 3. 音節数

	ドラえもん	のび太	スネ夫
新 / 旧	543 / 544	575 / 642	536 / 619

4.1 明瞭度・音質評価

Mean Opinion Score(MOS) 実験により，変換音声がアニメに適した音声であるかを調査するため，変換音声の発話内容の明瞭度と音質の劣化について評価した．被験者は明瞭度を，1 が全くわからない，5 がわかる，音質を，1 が非常に気になる，5 がわからないとする 5 段階で評価した (表 4)．

表 4. MOS 評価カテゴリ

評点	明瞭度	音質 (劣化)
5	わかる	わからない
4	だいたいわかる	わかるが気にならない
3	少しわからないところがある	やや気になる
2	わからない	気になる
1	全くわからない	非常に気になる

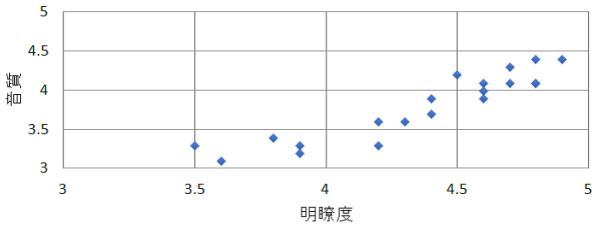

図 3. MOS 評価結果 (ドラえもん)

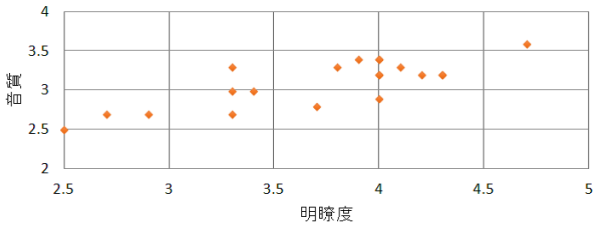

図 4. MOS 評価結果 (のび太)

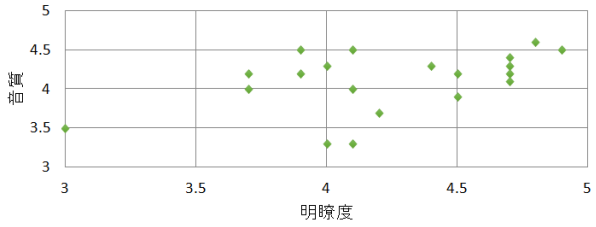

図 5. MOS 評価結果 (スネ夫)

評価音声データは，1 キャラクターにつき，ランダムに選択した 20 文であった．

図 3，4，5 に変換音声の発話内容の明瞭度，音質の結果を示す．ドラえもんは明瞭度が 4.4，音質が 3.8，のび太は 3.7，3.1，スネ夫は 4.2，4.1 であった．MOS が 3.5 以上であれば許容可能であり [7]，変換音声の発話内容の明瞭度はどれも 3.5 以上を示し，変換音声の音質は，のび太以外は 3.5 以上を示していた．明瞭度評価の低かったサンプルの特徴として，音質が悪いものが多かった．明瞭度と音質の相関関係を調査したところ，ドラえもんでは相関係数が 0.92，のび太では相関係数 0.80 と相関が見られた．これより，変換音声の発話内容の明瞭度と音質劣化は関係がある．

また，音質評価の低かったサンプルの特徴として，発話データ中の子音部分が主に音質劣化が感じられる傾向にあった．変換モデルは子音に比べ，母音の学習がはるかに多い．そのため，母音部分と子音部分で変換精度に差が出てしまった．これが音質劣化の原因となり，発話内容の明瞭度の低下の原因にもなったと考えられる．

4.2 話者性評価

音声 X が，音声 A と音声 B のどちらに近いかという ABX 評価により，変換音声が目標話者に近づいたかどうかを調査するため，変換音声の話者性について評価を行った．A，B には新声優，旧声優の自然音声を、X には変換音声を提示した．順序効果を抑えるため，A，B の順序をランダムにした．F_0 の違いによる偏りをなくすため，A，B，X のデータの F_0 の平均を求め，それぞれ昇順に 1 セットとした．評価音声データは話者性評価と同様とした．結果は，ドラえもん，のび太，スネ夫でそれぞれ 31.0% ，60.5% ，50.5% であった．変換の評定が良いとは言いがたい結果となった．

評定が悪い原因として，ドラえもんやスネ夫は，新旧の声優でそれぞれ異なったダミ声のような独特の発声をしていた．ダミ声はノイズによって調波成分の損失した発声 [10] で声帯振動に関係しており，今回は声道特性のみに着目した変換のため，

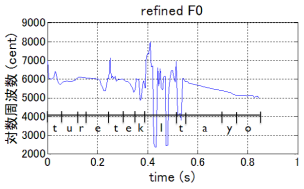

図 6. 旧声優

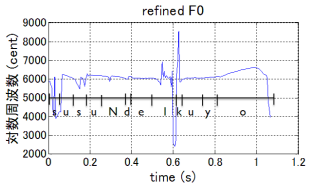

図 7. 新声優

評定の悪い原因の 1 つと考えられる．また，語尾の言い回し (伸ばし方など) やイントネーションなどの言葉遣いが特徴的な入力データも，変換後に特徴的な言葉遣いの印象が強く残った．(図 6，7) 新声優の自然音声が特徴の強く出ているデータであった場合，新声優と判断されており，特徴的な言葉遣いと判断誤りは相関があると考えられる。これは，特徴的な言葉遣いが特定の人物像を想起させる役割語 [8] となり，新声優が役割語の使用によって繰り出される発話キャラクタ [9] となったと考えられる。

さらに，パラレルコーパスの完成度も原因の 1 つであると考えられる．声優は発声のプロであり，一般人とは発声のスキル (発声方法，感情表現・伝達，声質，声量など) が異なるため，アライメントがうまくとれない可能性がある．また得られるデータに限りがあるため，より音韻バランスのよい学習データを用意できれば，変換モデルの完成度向上が期待できると考えられる．

5 あとがき

本研究では、同じキャラクターを演じる声優の声質変換を目的として，中間話者パラレルコーパスを用いた新声優から旧声優への声質変換手法を提案した．新声優と旧声優の同一発話内容の入出力音声対からなるパラレルコーパスの用意が困難であるため，2 者間に発話内容を自由に制御可能な中間話者を経由することで，声質変換を行った．変換モデル構築の際に音韻バランスを考慮したことで，少ないデータ数で変換モデルを構築できたものと考えられる．主観評価の結果，今回は声道特性のみに着目して変換を行ったが，声帯振動や F_0，イントネーションなど声道特性以外の特徴量の違いが見られた．しかし韻律的特徴は，感情の表現方法で特徴パラメータを適宜変化させているため，定量化が難しい．そこで，声帯振動を考慮すること，パラレルコーパス・変換モデルの完成度を向上させることで，話者性，発話明瞭度や音質がより良い結果を期待できると考えられ，今後の課題とする．

参考文献

[1] 伊藤, 他 ”音声の音響的特徴パラメータが個人性の知覚に及ぼす影響”, 信学論, Vol.J65-A, No.1, pp101-108, 1982.

[2] A.Kain, et. al. ”SPECTRAL VOICE CONVERSION FOR TEXT-TO-SPEECH SYNTHESIS,” Proc.ICASSP, p.285-p.288, May 1998.

[3] Tomoki Toda, et. al. ”Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory”, IEEE Trans. AUDIO, Speech, Lang, Process., VOL.25, NO.8, pp.2222-2235, 2007

[4] 磯, 他 ”音声データベース用文セットの設計”, 日本音響学会講演論文集, p.89-p.90, 1988.

[5] 匂坂, 他 ”ATR 音声・言語データベース”, 音響誌.48(12), p.878-p.882, 1992.

[6] 板橋修一, ”音声工学”, 森北出版, 2005, p.191-p197.

[7] 浅谷耕一, ”通信ネットワークの品質設計”, 信学会, 1993.

[8] 金水敏, ”ヴァーチャル日本語 役割語の謎”, 岩波書店, 2003.

[9] 定延利之, ”ことばと発話キャラクタ”, 文学, 2006, p.117-p129.

[10] Eiji Yumoto, et. al. ”Harmonics-to-noise ratio as an index of the degree of hoarseness”, J.Acoust.Soc.Am, Vol.71, NO.6, pp.1544-1550, 1982