韻律変換実現のための一試行：高橋みなみ風の音声を小嶋陽菜風に変えてみた

大野 涼平[1,a)]　北原 鉄朗[1,b)]

概要：本研究では，任意の音声の韻律をユーザが指定した音声の韻律に似せる方法の実現を目指している．これまで声質を変換する研究は数多くなされてきたが，韻律の変換については，あまり研究が行われていない．我々は，韻律変換実現のための第一歩として，韻律の特徴が大きく異なる小嶋陽菜と高橋みなみ (ともに尾木プロ) を取り上げ，次のことを試みた．1) アマチュア声優に同一文 (ATR 音素バランス文) を，小嶋陽菜および高橋みなみの韻律にできるだけ似せて発話してもらう (以下，「小嶋風音声」「高橋風音声」と呼ぶ)．2) 韻律のうち，基本周波数 (F_0) と話速が重要であると考え，音節ごとに F_0(または F_0 の変化量) と話速が小嶋風音声と等しくなるよう，高橋風音声の F_0 と話速を変換する．3) 被験者に変換後の音声を聞いてもらい，評価する．この結果，F_0 と話速を小嶋風音声と等しくすることで，変換なしの音声に比べて有意に小嶋らしくなることが分かった．

A Trial Towards Prosody Conversion: We Tried to Convert Minami Takahashi-like Speech to Haruna Kojima-like Speech

Ohno Ryouhei[1,a)]　Kitahara Tetsuro[1,b)]

Abstract: The goal of our study is to develop a system that converts the prosody of a given speech to that of a different speech. There have been many studies for voice conversion but only a little studies for prosody conversion. In this paper, as the first step, we tried to convert the prosody (specifically fundamental frequency and speech speed) of Minami Takahashi-like speech to that of Haruna Kojima-like speech. Experimental results show that, the generated speeches were significantly similar to Kojima's speech compared to the original speeches.

1. はじめに

我々は歌を聞いているときにその歌手の歌声や，テレビ番組を視聴しているときには出演者の声などに聴き惚れることがある．もしもユーザが任意の声を自分の好きな歌手やアイドルに似せた声に変換できれば新たなエンターテインメントの創出に繋がると期待される．話者変換に関する研究として主に統計的モデルに基づく声質変換の研究が盛んに行われている．入出力話者のパラレルデータの作成コスト削減のための手法 [1]，中間話者コーパスを用いた声優の声質変換 [2] などが提案されてきた．しかし声質のみならず韻律も話者性を与える大きな要因である．例えばモノマネなどで声質が酷似しておらずとも似ているという印象を受けるのは韻律が模倣できているかである．

韻律情報を自由に制御する手法として，入力音声の韻律を手本の音声のものに変換する声質変換 [3] や，HMM 音声合成において韻律パラメータにも話者適応を施したもの [4]，HMM 音声合成で出力する合成音声に対し，入力音声の韻律を模倣させるシステム [5] がある．これらの多くの研究では入出力話者の音声データ対のセット (以下，音声データセットと呼ぶ) の作成が必要になるが，同一発話文でなくてはいけないなど制約が多い．入出力話者がタレントなどの場合，制約を満たした音声データセットを十分に用意するのが困難である．

本研究では，対象話者をアイドルに限定し，同一発話文

[1] 日本大学 Nihon University
[a)] ryouhei@kthrlab.jp
[b)] kitahara@kthrlab.jp

の音声データセットを必要としない韻律変換システムの実現を目指す．具体的に，アイドルらしい韻律として「メイド声」[6] や「萌え声」[7] と呼ばれてるものに着目し，このような声が持つ韻律の特徴量を分析・再現する実験を行う．このような特徴を行う持つアイドルとして小嶋陽菜 (プロダクション尾木) を取り上げ，目標話者とする．一方，上記の特徴とは対称的な話し方をする高橋みなみ (プロダクション尾木) を入力話者とする．本稿では，入力音声と目標音声の韻律以外 (発話内容，声質) を同じにするため，アマチュア声優に同一文を小嶋風と高橋風の話し方で発話してもらうことで音声データセットを作成し，これを用いて高橋風の音声を小嶋風に変化させる試行を行う．その際，変換の度合いを 7 段階で制御し，より小嶋風に聴こえるものを調査する．

2. 音声データの作成と変換

本節では，実験に使った音声データセットの作成手順と変換処理の手続きについて述べる．変換処理においては，できるだけシンプルな処理でどこまで小嶋風の韻律を再現できるかを確かめるため，音節ごとの基本周波数 (F_0) と話速の時間軸方向の平均値のみを扱うこととする．

2.1 音声データの作成

本研究では，小嶋風の音声と高橋風の音声の韻律を比較するが，韻律のみを比較するには韻律以外の特徴 (発話内容，声質) は同じであることが望ましい．しかし，小嶋，高橋本人の音声でこれを実現するのは困難である．そこで，あるアマチュア声優の女性 (20 代) に小嶋，高橋の韻律を真似て同一文を発話してもらうことで，音声データ対を作成した．まず，事前に小嶋，高橋の音声を聴いてもらった．音声は，ラジオ番組「週刊ノースリー部」(ニッポン放送) から抜粋した約 681 秒の音声 (小嶋 425 秒，高橋 256 秒) である．その後，音素バランス文データセット [8] の A セット 01 から 10(表 1) をそれぞれ小嶋風，高橋風で発話してもらい，収録した．収録は発話者本人が納得いくまで繰り返すこととし，1 つの音声あたり平均 1.93 回の録音を行った．

表 1 ATR 音素バランス文 A セット

01	あらゆる 現実を すべて 自分の ほうへ ねじ曲げたのだ.
02	一週間ばかり ニューヨークを 取材した.
03	テレビゲームや パソコンで ゲームを して 遊ぶ.
04	物価の 変動を 考慮して 給水水準を 決める 必要がある.
05	救急車が 十分に 動けず 救急作業が 遅れている.
06	言論の 自由は 一歩 譲れば 百歩も 千歩も 攻め込まれる.
07	会場の 周辺には 原宿駅や 代々木駅も あるし ちょっと 歩けば 新宿御苑駅も ある.
08	老人ホームの 場合は 健康器具や ひざ掛けだ.
09	ちょっと 遅い 昼食を とるため ファミリーレストランに 入ったのです.
10	嬉しいはずが ゆっくり 寝ても いられない.

その次に，この女性に発声してもらった音声のそれぞれに，どの程度小嶋らしさ，高橋らしさが現れているかを評価し，これらが特に現れていると思われる 3 つの音声対を使用した．評価は男子大学生 6 名 (3 名は小嶋も高橋も顔と名前が一致しない，3 名は一致する程度の認知) にしてもらった．評価方法は次の通りである．

(1) 小嶋陽菜と高橋みなみの特徴を知ってもらうため，アマチュア声優女性に聴かせたものと同じラジオ音声を聴かせる．
(2) 録音音声 (小嶋風と高橋風音声 20 文) をランダムに聴かせ，どちらの声に聴こえたか評価させる．
(3) 各文章で，小嶋風，高橋風ともに全員が正解した音声セットを決める．

実験の結果，全 10 セットの中で A03， A09， A10 の 3 つの文を扱うこととする．

2.2 強制アライメントによる音声の分割

音声を音節単位で扱うため，Julius[9] で強制アライメントを行い，その結果に基づいて音声を音節単位で分割する．

2.3 F_0 の変換

小嶋風の音声と高橋風の音声に対して，5ms ごとに F_0 推定を行う．F_0 推定には World[10] を用いる．

今，小嶋風の音声の F_0 を $\{F^{\mathrm{K}}_{\mathrm{Hz}}(t)\}$，高橋風の音声の F_0 を $\{F^{\mathrm{T}}_{\mathrm{Hz}}(t)\}$ とする．これらを次式により cent 単位に変換する [11].

$$F^{p}_{\mathrm{cent}}(t) = 1200 \times \log_2 \frac{F^{p}_{\mathrm{Hz}}}{REF_{\mathrm{Hz}}} \quad (p = \mathrm{K}, \mathrm{T}) \tag{1}$$

$$REF_{\mathrm{Hz}} = 440 \times 2^{\frac{3}{12}-5} \tag{2}$$

次に，音節ごとの F_0 の平均値を計算する．i 番目の音節が始まる時刻を t_i，終わる時刻を t'_i とし，次式により計算する．

$$\overline{F^{p}_{i}} = \underset{t_i \le t \le t'_i}{\mathrm{mean}} F^{p}_{\mathrm{cent}}(t) \quad (p = \mathrm{K}, \mathrm{T}) \tag{3}$$

ここで，mean は平均値を求める関数である．これを基に，$\overline{F^{\mathrm{T}}_{i}}$ ができるだけ $\overline{F^{\mathrm{K}}_{i}}$ に近くなるように $\{F^{\mathrm{T}}_{\mathrm{cent}}(t)\}$ を変換し，再合成するが，変換については次の 2 つの方法を試行する．

2.3.1 方法 1

各音節の F_0 の平均値をそのまま使用する．高橋風の音声の F_0 を次のように変換する．

$$F^{\mathrm{new}}_{\mathrm{cent}}(t) = F^{\mathrm{T}}_{\mathrm{cent}}(t) + \alpha(\overline{F^{\mathrm{K}}_{i}} - \overline{F^{\mathrm{T}}_{i}}) \quad (t_i \le t \le t'_i) \tag{4}$$

ここで，α は変換の度合いを表す係数で，$\alpha = 1$ のときは $\overline{F^{\mathrm{new}}_{i}} = \overline{F^{\mathrm{K}}_{i}}$ でる．$\alpha = 0$ のときは変換は行われず，$\overline{F^{\mathrm{new}}_{i}} = \overline{F^{\mathrm{T}}_{i}}$ である．$\alpha > 1$ にすることで，小嶋の特徴を「おおげさ」に再現することができる．

2.3.2 方法 2

隣接する音節の F_0 の平均値の差を使用する．変換後の音声における隣接する音節の F_0 の平均値 $\overline{F_i^{\mathrm{new}}} - \overline{F_{i-1}^{\mathrm{new}}}$ $(i > 1)$ が，小嶋風の音声のそれ $\overline{F_i^{\mathrm{K}}} - \overline{F_{i-1}^{\mathrm{K}}}$ と等しくなるようにしたい ($\alpha = 1$ のとき)．それで，F_0 の変換を 1 番目の音節から順番に行うものとし，

$$F_{\mathrm{cent}}^{\mathrm{new}}(t) = F_{\mathrm{cent}}^{\mathrm{T}}(t) - \overline{F_i^{\mathrm{T}}} + \alpha(\overline{F_i^{\mathrm{K}}} - \overline{F_{i-1}^{\mathrm{K}}}) + \overline{F_{i-1}^{\mathrm{new}}} \quad (5)$$

とする．なお，$i = 1$ のときは F_0 の変換は行わない ($F_{\mathrm{cent}}^{\mathrm{new}}(t) = F_{\mathrm{cent}}^{\mathrm{K}}(t)$) ものとする．

2.4 話速の変換

2.2.1 節と 2.2.2 節の方法で F_0 を変換した高橋風の音声に対して，音節ごとの話速が小嶋風の音声のものと同じになるように音響信号を時間軸方向に伸縮させる．

i 番目の音節における話速を S_i^p $(p = \mathrm{K}, \mathrm{T})$ とすると，変換後の話速は次のようになる．

$$S_i^{\mathrm{new}} = S_i^{\mathrm{T}} \left(\frac{S_i^{\mathrm{K}}}{S_i^{\mathrm{T}}} \right)^{\alpha} \quad (6)$$

音響信号の伸縮には，それによって音の高さが変わらないよう，Phase Vocoder(Ellis が実装したもの [12]) を用いる．ここで，α は 2.3.1 節および 2.3.2 節と同じ効果である．

3. 評価実験

3.1 実験方法

2 章で述べた手法を用いて変換した音声が，聴感上どの程度小嶋の韻律に近づいているか，被験者実験によって確かめた．まず，α を 0 から 1.5 まで 0.25 ずつ変えて音声の変換を行った．原理的には α の値は F_0 と話速と別の値を割り当てることができるが，評価対象の音声が増えるのを避けるため，同じ値を用いることとした．このようにして用意した 7 個の音声の全てのペア (21 組) に対して，どちらの方が小嶋らしいかを次の 5 段階で評価してもらった:

- 1 つ目の方が小嶋らしい (4)
- 1 つ目の方がやや小嶋らしい (2)
- どちらとも言えない (0)
- 2 つ目の方がやや小嶋らしい (-2)
- 2 つ目の方が小嶋らしい (-4)

括弧内の数字は，後述の分析時に用いる評価値である．

評価にあたっては，被験者には予め「小嶋風の音声」と「高橋風の音声」(両方とも本人の音声ではなく，当該アマチュア声優が似せて発声した音声) を聴いてもらってから評価してもらった．評価は，同じ音声に対して 2 回ずつ行い，F_0 変換手法として方法 1 を用いた場合と方法 2 を用いた場合とでそれぞれ行った．評価対象の音声は 2.1 節で選んだ 3 つの音声 (A03, A09, A10) である．被験者は，20 代男性 5 名 (3 名は声を聴いたときにそれが小嶋，高橋のものだとわかる程度に認知している) で行った．

図 1 実験により得られた α の値と平均評価値の関係 (F_0 変換は方法 1 を利用)

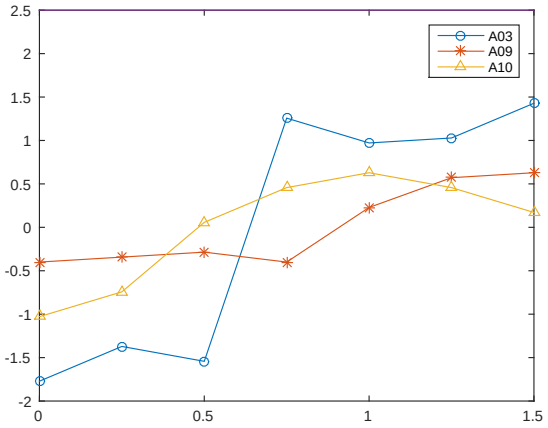

図 2 実験により得られた α の値と平均評価値の関係 (F_0 変換は方法 2 を利用)

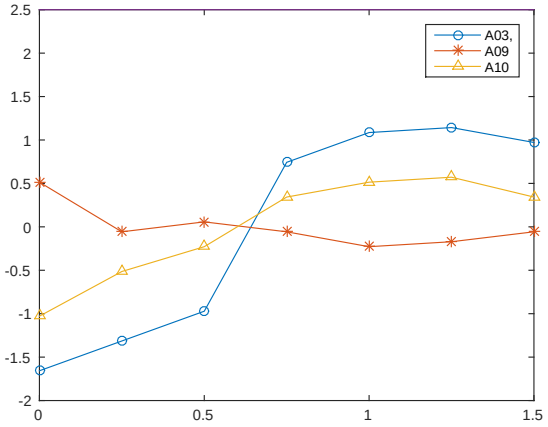

表 2 変換後の音声対の平均評価値に対して有意差の有無を決める閾値：yardstick (F_0 の変換は方法 1 を利用)

	$Y_{0.05}$	$Y_{0.01}$
A03	0.7976	0.94688
A09	1.0127	1.2022
A10	0.9842	1.1684

表 3 変換後の音声対の平均評価値に対して有意差の有無を決める閾値：yardstick (F_0 の変換は方法 2 を利用)

	$Y_{0.05}$	$Y_{0.01}$
A03	0.9398	1.1156
A09	0.9074	1.0772
A10	0.8623	1.0236

3.2 実験結果

評価結果について，シェッフェの一対比較法に対する中屋の変法 [13] を用いて分析を行った．一対比較を行う 21 組の各ペアの平均評価値の差が表 2，3 の該当する値を超えていたら有意差があるものと判断する．例えば，方法 1 において A03 から $\alpha = 0.5$ と $\alpha = 1.25$ の差は 2.5714 であった．これは表 2 で $Y_{0.01} = 0.94688$ を超えている．すなわ

ち危険率 1% で評価対象間に有意な差があると言える．

これにより，各音声がどの程度小嶋らしいかを推定することができる (推定値をここでは「小嶋らしさ」と呼ぶ)．各音声の変換時に指定した α の値を横軸，小嶋らしさを縦軸として結果を表したものを図 1，図 2 に示す．

F_0 に方法 1 を用いた場合 (図 1) では，α の値が大きいと「小嶋らしさ」も向上するという結果になっている．実際，A03 は $\alpha = 0$ と $\alpha \geq 0.75$，$\alpha = 0.25$ と $\alpha \geq 0.75$，$\alpha = 0.5$ と $\alpha \geq 0.75$ で危険率 1% で有意な差が見られた．A09 は $\alpha = 0$ と $\alpha = 1.5$，$\alpha = 0.75$ と $\alpha = 1.5$ で危険率 5% で有意な差が見られた．A10 は $\alpha = 0$ と $\alpha = 0.5$ の間で危険率 5% で有意な差が，$\alpha = 0$ と $\alpha \geq 0.75$，$\alpha = 0.25$ と $0.75 \leq \alpha \leq 1.25$ の間で危険率 1% で有意な差が見られた．このように，変換によって，有意に小嶋らしさを向上させることができた．これは，小嶋は高橋に比べ全体的に F_0 が高い，F_0 の音節間の変化が大きい，話速がゆっくりであるという特徴があるが，これがより強調されたからと考えられる．ただし，A10 では $\alpha = 1.0$ が最も小嶋らしさが高くなり，α を上げると小嶋らしさが下がるという結果が得られた．これは，変換によって極端に話速が低くなる箇所があり，/r/の音が伸ばされたために巻き舌のような音になり，これにより小嶋らしさが低下したからではないかと考えられる．F_0 が大きく変わる部分における音声の品質劣化も見られ，こちらも影響した可能性もある．

F_0 変換に方法 2 を用いた場合 (図 2) では，A03 は $\alpha \leq 0.5$ と $\alpha \geq 0.75$ の間で危険率が 1% で有意な差が見られ，A10 では $\alpha = 0$ と $\alpha \geq 0.75$，$\alpha = 0.25$ と $1.0 \leq \alpha \leq 1.25$ で危険率 1% で有意な差が見られた一方，A09 では有意な差は見られず，A09 ではむしろ小嶋らしさが $\alpha = 0$ と比べて低下する結果となった．これは，A09 の高橋風音声には小嶋風音声よりも話速が低い部分が含まれており，変換によって話速が上がってしまうので，小嶋らしく聴こえなくなったからと考えられる．

4. おわりに

本稿では，任意の音声の韻律をアイドルらしく変換するシステムの実現を目指し，高橋みなみ風に発声さらに音声の韻律を小嶋陽菜風の音声の韻律に近付ける実験を行った．音節ごとの F_0 と話速を変えるだけでも一定程度小嶋らしく聴こえるようになることがわかった．一方，極端に話速や F_0 が変化する場合には，音声の劣化によってむしろ小嶋らしく聴こえなくなることもわかった．

今回の実験では，同一文を発話した高橋風の音声と小嶋風の音声を実験に用いたが，実用を考えた場合，入力音声と目標音声の発話文が同一であることは期待できない．今後は，音声変換の品質はもとより，発話内容に依存しない特徴抽出と特徴変換の検討が必要となる．

謝辞 本研究では，評価実験や音声データ作成でご協力下さった全ての皆様に感謝いたします．

参考文献

[1] 中鹿 亘, 滝口 哲也, 有木 康雄: “話者適応型 Restricted Boltzmann Machine を用いた声質変換の検討”, SP2014-126, pp.165-170, 2014.

[2] 塩出 萌子, 小泉 悠馬, 伊藤 克亘: “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”, 情報処理学会論文誌 (第 76 回全国大会), vol.1, pp.495-496, 2014-3.

[3] 足立 吉弘，森島 茂生: “話者のイントネーションを模倣するインタラクティブ声質変換システムの構築”, 情報処理学会シンポジウム論文集, 2005-4, pp.261-268,2005-2.

[4] 田村 正統, 益子 貴史, 徳田 恵一, 小林 隆夫: “HMM に基づく音声合成におけるピッチ・スペクトルの話者適応”, 電子情報通信学会論文誌, D-Ⅱ, Vol.J85-D-Ⅱ, No.4, pp.545-553, 2002-4.

[5] 西垣 有理, 高道 慎之介, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲: “音声入力による韻律制御機能を有する HMM 音声合成システムの改良”, 2014-SLP-104(16), 1-6, 2014-12.

[6] Kawahara,Shigeto: “The phonetics of Japanese maid voice l: vA preliminary study”, Phonological Studies, 16, pp.19-28, 2013.

[7] 高野 佐代子, 竹澤 勇希, 竹内 純基, 山田 真司: “「萌え声」心理的評価，音響分析および STRAIGHT を用いた合成音声評価”, 日本音響学会 2014 年春季研究発表会聴講論文集, No.2-Q5-22, 2014.

[8] 匂坂 芳典, 浦谷 則好: “ATR 音声・言語データベース”, 日本音響学会誌, 48(12), pp.878-882, 1992.

[9] 李 晃伸: “大語彙連続音声認識エンジン Julius ver. 4”, 情報処理学会研究報告. SLP, 音声言語情報処理 69, pp.307-312, 2007-12-20.

[10] 森勢 将雅，西浦敬信，河原 英紀: “高品質音声分析変換合成システム WORLD の提案と基礎的評価　基本周波数・スペクトル包絡が品質の知覚に与える影響 ”, 日本音響学会, vol.41, No.7, pp.555-560, Toyama, Oct.1-2, 2011.

[11] 後藤 真孝: “実世界の音楽音響信号を対象としたメロディーとベースの音高推定”, 99-MUS-31-16, Vol.99, No.68, 1999-8.

[12] D.P.W.Ellis: “A Phase Vocoder in Matlab”, http://www.ee.columbia.edu/ dpwe/resources/matlab/pvoc/, 2002.

[13] 高木 英行: “使える！統計検定・機械学習ー Ⅱー主観評価実験のための有意差検定”, システム制御情報学会誌 58(12), 514-520, 2014-12-15.