平成 27 年度 卒業論文

かわいらしい音声への加工のための韻律分析

指導教員 北原鉄朗准教授

日本大学文理学部情報システム解析学科

大野 涼平

2016 年 2 月 提出

概　　要

アニメは日本のサブカルチャーで最も人気なものの 1 つである．特に女性声優がかわいらしさを強調した声，いわゆる「萌え声」は注目を浴びており，アニメ人気に大きく貢献している．従来，入力音声をユーザ任意の音声に変換する音声変換の研究は多く為されてきたが,「かわいらしい」音声など特定のテーマに特化した音声へ変換する研究は十分に為されていない．しかし，実際に入力音声をユーザ任意の「かわいらしい音声」に変換できるシステムが実現できれば，動画制作やゲーム制作などの新たなエンターテインメントの創出に期待できる．

本研究では，それを実現する第 1 段階として，同一話者の同一発話内容における地声と「かわいらしい音声」の基本周波数変化パターンのどこに違いが出るかを分析し，地声をどの程度までかわいらしく出来るかを確かめる．また本研究では,「かわいらしい音声」を，女性がかわいさらしさを強調した声と定義する．

基本周波数（F0）変化パターンの分析結果では，地声とかわいらしい声の F0 平均値の差は 222 [cent] 程度であり，平均値を合わせても尚，両音声間の文節単位での F0 極大値の差に 139.1 [cent] の有意な差があるということがわかった．

また，実際に地声を入力とし，かわいらしい声を目指し変換した音声を聴取評価した結果，F0 成分の高さを変えるだけではかわいらしい声になるかはわからないという結果になった．

目　　次

目　　次 **iii**

図 目 次 **vii**

表 目 次 **ix**

第 1 章　序　　論 **1**

1.1　本研究の背景 1

1.2　本研究の目的 2

1.3　本論文の構成 2

第 2 章　関連研究 **3**

2.1　話者変換 3

2.1.1　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 [1] 3

2.1.2　音声の韻律情報の変換によるイントネーション変換システム [2] 4

2.1.3　参照話者を用いた多対多固有声変換法 [3] 4

2.1.4　話者適応型 Restricted Boltzmann Machine を用いた声質変換の検討 [4] 4

2.2　韻律分析 5

2.2.1 「萌え声」心理的評価、音響分析及びSTRAIGHTを用いた合成音声評価 [11] 5
2.2.2 The phonetics of Japanese maid voice I: A preliminary study [12] 5
2.2.3 音声対話システムのための親近感特徴量の探索 [15] 6
2.3 関連研究のまとめ 6

第3章 データベースの作成と分析方法 7
3.1 音声データの取得 7
3.2 データの選別とラベル付け 8
3.3 分析方法 8
3.3.1 F0平均値の分析 9
3.3.2 F0極大値の分析 9

第4章 分析結果と考察 13
4.1 近似精度 13
4.2 F0平均値の分析 13
4.3 F0極大値の分析 14

第5章 音声変換への応用 17
5.1 使用データ 17
5.2 変換手順 17
5.3 変換精度と考察 18
5.3.1 F0平均値の客観評価 18
5.3.2 F0極大値の客観評価 20
5.4 主観評価 22

第6章 結論 31

参考文献 **33**

図 目 次

3.1 1 文におけるオリジナル F0 の軌跡 10
3.2 1 文におけるオリジナル F0 と近似 F0 の軌跡（文節毎に分割した）
(/ペットボトルで/ /何を/ /するかと/ /思ったら、/ /ペットボトル
ロケットだったんだねぇ。/は，B-S-S-Sc-E である．) 10

5.1 F0 変換の流れ 19
5.2 上段：2 番目の音声　中段：15 番目の音声　下段：21 番目の音声　. . 27
5.3 3 番目のセットにおける 3 種類の音声の F0
(ねぇねぇ、さっきの 授業、何 やってたの？) 28
5.4 18 番目のセットにおける 3 種類の音声の F0
(でもさぁ、もう 夜中の 12 時なのに、相変わらずの 夕日なんだね。) 29
5.5 13 番目のセットにおける 3 種類の音声の F0
(裕太くんったら 私が いないと ダメんだんだからぁ) 30

表 目 次

3.1 文節のラベル 8

4.1 ユークリッド距離 13
4.2 F0 平均値の差 14
4.3 F0 極大値の差 15

5.1 平均値の差の検定結果 20
5.2 極大値の差の検定 21
5.3 ラベル毎による極大値の差の検定 22
5.4 設問 1 22
5.5 設問 2 22
5.6 設問 1 の評価値・改 23
5.7 設問 1 の評価平均値と標準偏差 23
5.8 設問 2 の評価平均値と標準偏差 24
5.9 目標・出力と出力・入力音声のユークリッド距離 26

第1章　序　　論

研究の背景，目的，従来研究との違いなどを，過去の論文を引用しながら述べる．

1.1　本研究の背景

近年，音声信号処理の分野において，音声合成（コンピュータ上で自然な音声を生成する技術）に関する研究が盛んである．単に自然な音声を合成するだけでなく魅力的な声や特徴的な声の実現への注目が特に高まっている．その例として声質変換 [1, 2, 3, 4] や，テキスト情報から自然な音声を生成するテキスト音声合成 [5, 6] などがある．また，日本の最も人気なサブカルチャーの 1 つであるアニメにおいて，その女性声優がかわいらしさをより強調した音声は「萌え声」とも言われ特に高い人気を誇っており，近年では女性声優がアイドル化するなど女性声優にまで注目が集まっている．さらに音声合成ソフトウェアである Vocaloid や Cevio では若い女性の声を使用しており，ここでも「かわいらしい声」が人気を高めていることがわかる．声質変換やテキスト音声合成技術を使うことで，かわいらしい音声など特徴的な音声の合成がある程度実現できる．しかし，喋り方（韻律）も話者の個人性や話者の持つ印象を表現しているということがわかっており [7, 8, 9, 13]，上記の技術だけでは目標とする声の韻律を再現できないという問題が残る．

しかし「かわらしい」の基準は個人によって不定であり,「かわいらしい声」を定義付けることは困難である．従って「かわいらしい声」が人気であってもユーザ任意の「かわいらしい声」を合成する技術は提案されておらず,「かわいらしい

声」は一体どのような声なのかという分析に留まっており，尚もその正体は解明されていない [10].

1.2 本研究の目的

本研究では，入力した地声をユーザ任意のかわいらしい声に変換するシステムの実現を目指し，その第一段階として同一話者の同一発話文による地声とかわいらしい声とのの間に現れる韻律情報の差，特に音高を表す基本周波数（F0）の変化パターンの差を分析する.

また，それを実際に音声変換へと応用したときにどの程度までかわいさを模倣することが可能かを調べる.

1.3 本論文の構成

本論文の構成を述べる．2章では本研究の関連研究について述べる．3章では本研究で使用する音声データベースの作成手順，音声データの前処理，分析方法について述べる．4章では分析結果と考察を述べ，5章では4章の結果に基づき音声変換への応用について述べる．6章では本研究のまとめの今後の展望についてを述べる.

第2章　関連研究

本章では今まで行われてきた関連研究について，話者変換，韻律分析の観点からいくつか述べる．

2.1　話者変換

音声の音響特徴量を他の話者のものへと変換する技術の実現や応用に関する研究については，既に様々なものが存在する [1, 2, 3, 4, 5, 6]．以下では，代表的なもののみ述べる．

2.1.1　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 [1]

従来，統計的声質変換（話者変換）では元話者（入力話者）と目標話者（出力話者）それぞれが同一文を十分な量かつ明瞭に発話した音声データ対（パラレルデータ）を予め用意しておく必要がある．しかし，声優間の変換を行う場合，上記の条件を満たした音声データ対を用意することは困難であるために有名人同士の話者変換は実現できなかった．そこでこの研究では，元話者と中間話者，中間話者と目標話者それぞれでパラレルデータを作成して変換モデルを作成する．この中間話者を設けることでデータ対の取得が困難である話者間での話者変換が可能となった．

2.1.2　音声の韻律情報の変換によるイントネーション変換システム [2]

予め用意した入力話者の音声の韻律情報を参照話者のものへと変換することで，入力話者の声質を参照話者のものへと変換可能となる．また，このシステムは前述の条件を満たす音声データ対を必要とせず，変換したい音声データがあるだけで十分である．入力音声の韻律情報を，話速，F0，パワーの順に参照話者のものへと変換していく．同時に，話速変換時に線形補間，Catmull-Rom 補間，3 次スプライン補間のどれを使用すると劣化なく変換できるかを調査しており，どれも差がないと報告がある．以上により，収録し直すことが不可能な音声の韻律を自由に制御することが可能となった．

2.1.3　参照話者を用いた多対多固有声変換法 [3]

固有声声質変換とは、予め用意した大多数の話者間におけるパラレルデータを用意し，固有声 GMM を学習する．固有声 GMM の固有ベクトルに対する重みを少量制御することで，声質を制御することが可能となった．

2.1.4　話者適応型 Restricted Boltzmann Machine を用いた声質変換の検討 [4]

入出力話者のパラレルデータのみならず参照話者間のパラレルデータさえ不要とする声質変換システムを目指している．適応型 RBM は，参照話者のデータで話者依存重みと話者非依存重みを推定し，元話者のデータで話者非依存重みを固定しながら話者依存重みを推定する．また目標話者のデータの重みを同様に推定する．次に，入力音声から元話者の重みを推定し，目標話者の重みを用いて音響特

徴量を逆推定することで変換する．低域スペクトルにおいては異性間の変換もパラレルデータを使用せずおおよそ一致している．

2.2 韻律分析

対象話者の喋り方（韻律）と比較話者の韻律，あるいはある話者の地声と演技声の韻律の間にどのような差が現れるかを分析した研究をいくつか述べる．

2.2.1 「萌え声」心理的評価、音響分析及びSTRAIGHTを用いた合成音声評価 [11]

「萌え声」は音声による感情表現ではなく話者の生態学的な情報であるとし，萌え声の設計指針を示した．分析対象音声は「お兄ちゃんCD」[14] から使用している．その結果，F0平均値，F0標準偏差，時間長の3点が萌え度に関係するとわかり，これらを独立に操作すると萌え度は向上させることができると報告している．

2.2.2 The phonetics of Japanese maid voice I: A preliminary study [12]

メイド喫茶のメイドによるメイド声とその人の地声の2つの違いを音響的に分析した．その結果，F0変化ではメイド声の方が全体的に高く，上昇部分においてその増加量が大きく，1秒毎の変化量が大きいということがわかった．また，全体的に声の強さはメイド声のの方が強く，フォルマント周波数にも大きな違いが現れている．

2.2.3　音声対話システムのための親近感特徴量の探索 [15]

友人同士と面識のない人同士，それぞれの組の 100 発話程度の音声を分析し，親近感が高い場合と低い場合の音響特徴量の差について調べた．その結果，親近感が低い場合は高い場合と比べ，F0 の変化量が低く，単調な発話になりやすいということがわかった．

2.3　関連研究のまとめ

話者変換 [1, 2, 3, 4, 5, 6] に関して，声質変換では話者が演技をすることで自由に「かわいらしさ」を制御できることから，声質の転写だけではかわいらしい声に変換するのは困難である．また，かわいらしい声の F0 変化パターンの定式化も困難であるため，任意の入力音声に対する韻律変換システムの実現も困難である．韻律分析では同一話者でも韻律特徴量を変えるだけでその話者の印象・特徴を別の話者のものに変えることが出来るとわかっている [7, 8, 9, 13]．特に F0 の大きさと変化量がかわいらしい声に近い意味を持つ「萌え声」や「メイド声」と関係があることがわかった [11, 12]．以上を踏まえ，かわいらしい声と地声の比較には F0 変化パターンの全体的な平均値及び変化量に着目することが望ましいとわかる．また，人の話し声の F0 パターンを扱う場合，藤崎モデル [16] がよく使われる．これは時間をかけてなだらかに低くなっていく成分（フレーズ成分）と，局所的に上下する成分（アクセント成分）の畳み込みからなるモデルである．しかし，音声から抽出した F0 から，各成分を推定することは逆問題であり，成分推定には複雑なアルゴリズムを要する必要がある．本研究ではいかに簡素なシステムでどれほどかわいらしい音声を再現できるか調べることを目的としているため，今回は用いない．

第3章　データベースの作成と分析方法

文献 [11, 12] ではかわいらしい音声と F0 平均値の高さに関する関係性について述べている。そこで本研究では、ある話者がかわいらしく演技したときの演技音声における F0 平均値、F0 ピークが、同一話者の地声のそれらより有意に高いという仮説を立てる。本章では，分析に使用するデータベースの作成方法と分析方法について述べる．

3.1　音声データの取得

分析には同一話者による同一発話文における地声とかわいらしい声の音声データ対が必要である．そこで、本研究では①両発話スタイルを使い分けることのできる女性話者，② 両発話スタイルで発話しやすい文　の 2 点を考慮しなければならない．従って，①の考慮として，声優経験のある女性（20 代）1 名に協力してもらった．②は，舞台台本である「夕映えワンダーホール」[17] の登場人物から女子高生の「朝日」役としてそのセリフを発話してもらう．発話の際に，協力者の地声，かわいらしさを強調した演技音声（以下，演技声）の 2 パターンでそれぞれ読み上げてもらい録音した．各パターンを 1 回ずつ，これを 3 日間にかけて計 3 回ずつ録音した．

表 3.1: 文節のラベル

B	文頭（下記以外）
Bc	文頭（読点で終わる）
S	文中（下記以外）
Sc	文中（読点で終わる）

E	文末（句点で終わる）
Eq	文末（疑問符で終わる）
Ea	文末（感嘆符で終わる）

3.2　データの選別とラベル付け

地声，演技声はそれぞれ 86 文，235 文節からなる．まず，両音声を Julius のセグメンテーションキット [18] でアライメントを取り分割する．アライメント誤りは著者が手動で修正した．1 回目，2 回目に録音した全音声の文節（235 × 4 文節）で，各文節毎に 4 つの音声をランダムに聴き，①最も演技声に聴こえた，　②2 番目に演技声に聴こえた，③2 番目に地声に聴こえた，　④最も地声に聴こえた，の 4 つに分類した．これは著者を含む男性 3 名でそれぞれ独立に行った．この分類結果から，各文節において①に分類した文節音声が 3 人中 2 以上が一致した場合，その文節における演技声を①の文節音声に，一致しなかった場合は②に定義した．地声も同様に定義した．また，定義した 235 文節に対し，表 3.1 に従ってラベルを付与する．

3.3　分析方法

まず，各文節の F0 を WORLD[19] によって 5ms 毎に抽出し，これを cent 単位に変換する．また，本研究では文献 [11, 12] を参考に，地声と演技声で文節単位の F0 平均値と極大値の差を分析する．

F0 は時間的微細に変動する (図 3.1)．この微細変動は発話内容を構成する単語などに影響を受けると考えられるため，用意した音声データのみではあらゆる発

話内容には対応できない．そこで単語などの言語情報に一般性を与えるためにこの微細変動を消し，F0 変化のおおまかな近似軌跡を分析するべきである．F0 変化方向は上昇，下降，平行のいずれかで，日本語は同一文節内でアクセント核を 1 つしか持たないことから，F0 を二乗誤差が最小となるよう 3 次曲線に近似して扱う．以降，F0 はこの 3 次曲線に近似したものとする．また，F0 パターンを「F0 モデルより本手法の方が良いという点を挙げたい」図 3.2 に例を示す．また，i 番目の文節における地声，演技声の F0 をそれぞれ $\{F_i^{\mathrm{N}}(t)\}$，$\{F_i^{\mathrm{A}}(t)\}$ とする．

3.3.1 F0 平均値の分析

文献 [11, 12] を参考に，本研究でも演技声の方が地声よりも F0 が全体的に有意に高いという仮説を立てる．i 番目の文節における演技声の F0 の時間平均を $\overline{F_i^{\mathrm{A}}} = \mathrm{mean}_t F_i^{\mathrm{A}}(t)$，地声のそれを $\overline{F_i^{\mathrm{N}}} = \mathrm{mean}_t F_i^{\mathrm{N}}(t)$ とする．$\overline{F_i^{\mathrm{A}}}$ が $\overline{F_i^{\mathrm{N}}}$ より有意に高い値を取るかを，表 3.1 のラベルごとに t 検定で調べる．また，有意な差が見られたラベルが複数個あった場合，それらのラベル間で $\overline{F_i^{\mathrm{A}}}$ と $\overline{F_i^{\mathrm{N}}}$ との差同士に有意な差があるかを再び t 検定で，それらラベルが 3 個以上の場合は多重比較によって調べる．これはラベル毎に話者が声の高さを制御しているのか，あるいは演技時には一律に声の高さを制御しているのかを調べるためである．

3.3.2 F0 極大値の分析

文献 [12] を参考に，平均値を上げた地声の F0 における極大値よりも演技声の F0 における極大値の方が有意に高いという仮説を立てる．ただし，両音声の F0 に極大値が存在しない文節は分析対象外とする．

ラベル L による平均値の差 α_L を 3.3.1 節の分析結果に従って定める．文節 i において地声と演技声の F0 が極大値を取る時刻をそれぞれ τ_i^{N}，τ_i^{A} とする．ラベル L における $\{F_i^{\mathrm{N}}(\tau_i^{\mathrm{N}}) + \alpha_L\}$ と $\{F_i^{\mathrm{A}}(\tau_i^{\mathrm{A}})\}$ に有意な差があるかを t 検定で調べる．ま

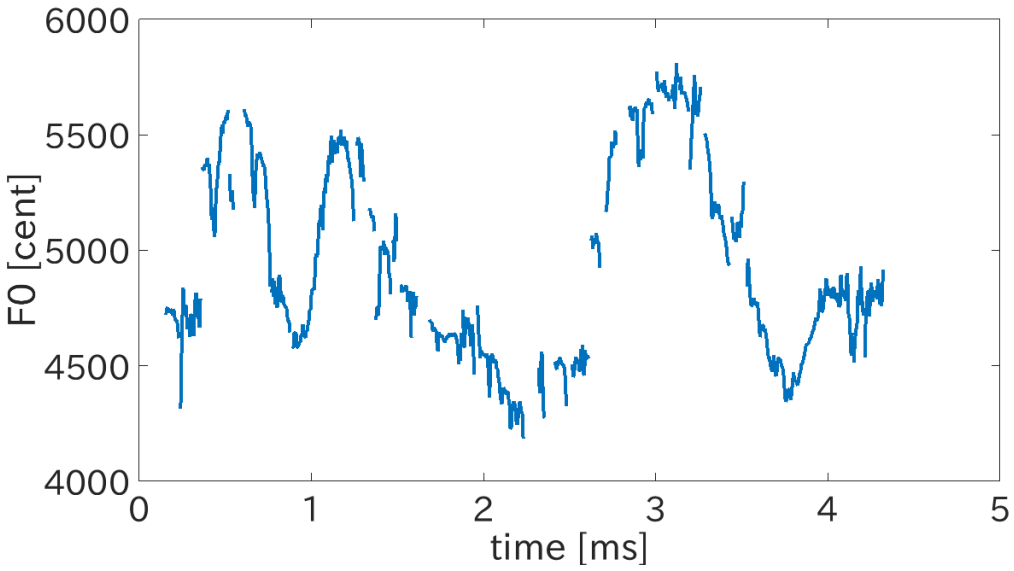

図 3.1: 1 文におけるオリジナル F0 の軌跡

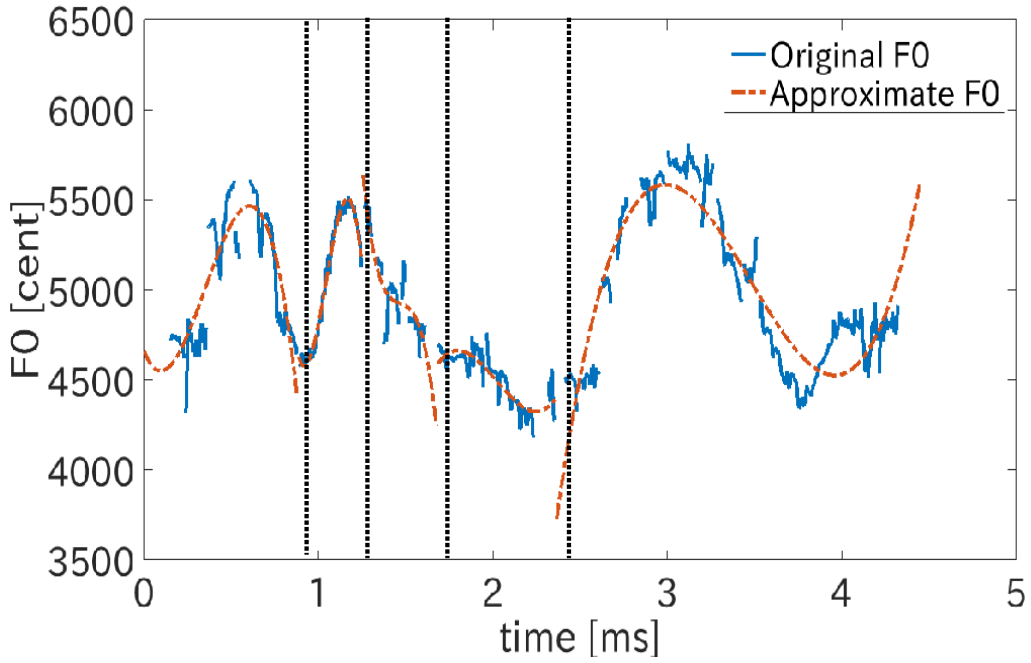

図 3.2: 1 文におけるオリジナル F0 と近似 F0 の軌跡（文節毎に分割した）
(/ペットボトルで/ /何を/ /するかと/ /思ったら、/ /ペットボトルロケットだったんだねぇ。/は，B-S-S-Sc-E である．)

た，有意差の見られるラベル間の差に有意差があるかを再び t 検定で，そのラベルが 3 個以上の場合は多重比較によって調べる．これはラベル毎に話者がアクセントを強めているのか，あるいは一律にアクセントを強めているのかを調べるためである．

第4章　分析結果と考察

本章では3章の分析結果とその考察について述べる．

4.1　近似精度

本研究ではcent単位のF0を3次曲線で近似して扱った．その近似精度として，オリジナルのF0と近似F0のユークリッド距離の平均値を算出した（表4.1）．これらから一定程度の精度で3次曲線に近似できていることがわかる．

4.2　F0平均値の分析

まず，全ラベルにおける $\{\overline{F_i^{\mathrm{A}}}\}$ と $\{\overline{F_i^{\mathrm{N}}}\}$ の差に t 検定を行ったところ，有意な差が見られた（$p = 2.37 \times 10^{-30}$）．ラベル毎に行った検定結果は表4.2のようになり，ラベルBc以外においては有意な差が見られ，Bcでは有意傾向が見られた．従って，文献[11, 12]と同じように，地声よりも演技声の方が全体のF0が有意に高いことが示された．また，全2つのラベル21組で $\overline{F_i^{\mathrm{A}}}$ と $\overline{F_i^{\mathrm{N}}}$ の差同士に有意な差があるかを多重比較で検定を行った結果，どの組にも有意な差があるとは言えなかっ

表 4.1: ユークリッド距離

	平均値	標準偏差
地声	0.10×10^{-11}	0.10×10^{-11}
演技声	0.09×10^{-11}	0.08×10^{-11}

表 4.2: F0 平均値の差

ラベル	平均値	標準偏差	p 値
B	176.7	225.3	0.0000
Bc	194.5	210.3	0.0004
S	194.4	196.2	0.0000
Sc	336.9	172.1	0.0000
E	314.9	266.7	0.0000
Eq	270.1	129.3	0.0000
Ea	211.4	170.9	0.0000
全ラベル	222.7	206.5	0.0000

た．すなわち，地声と演技声の全体の F0 は文節毎に有意な差を持つのではなく，全文節で一律に同程度の差を持つということである．全ラベルにおける $\{\overline{F_i^{\mathrm{A}}}\}$ と $\{\overline{F_i^{\mathrm{N}}}\}$ の差の平均値は 220.6 cent であった．

4.3　F0 極大値の分析

4.2 節に基づいて，$\alpha_L = 222.7$ [cent] と定める．

ラベル毎に $\{F_i^{\mathrm{A}}(\tau_i^{\mathrm{A}})\}$ と $\{F_i^{\mathrm{N}}(\tau_i^{\mathrm{N}}) + \alpha_L\}$ の差が有意な差であるかを t 検定にて調べたところ，表 4.3 の通りとなり，Sc，E のときに有意傾向あるいは有意な差がみられた．この 2 つの差に有意な差が有るか再び t 検定を行った結果，有意な差があるとは言えなかった．これは Sc，E における地声と演技声の F0 極大値は一律に同程度の差があるということである．Sc，E における $\{F_i^{\mathrm{A}}(\tau_i^{\mathrm{A}})\}$ と $\{F_i^{\mathrm{N}}(\tau_i^{\mathrm{N}}) + \alpha_L\}$ の差の平均値は 161.7 cent であった．

表 4.3: F0 極大値の差

ラベル	平均値	標準偏差	p 値
B	-21.6	224.6	0.7049
Bc	22.2	220.2	0.3462
S	18.9	208.4	0.7597
Sc	83.8	177.3	0.0971
E	161.8	195.2	0.0004
Eq	-63.7	237.7	0.8328
Ea	-64.5	198.3	0.9377
Sc,E	139.1	190.6	0.0002

第5章　音声変換への応用

第 4 章の分析結果を踏まえ，実際に地声を入力とし，かわいらしい声に変換する手法を提案する．変換音声の近似 F0 の変換精度を算出し，考察する．

5.1　使用データ

事前に収録した，分析時と同じ文を発話した 3 回目の録音音声（地声・演技声各 86 文ずつ）から，十分に地声と演技声に聴き分けられると判断できる 26 対を使用する．まず，被験者には 1 対の音声を 3 回ずつ聴いてもらう．片方は地声で，もう片方は演技音声である．このとき，どちらが地声だと思うか回答してもらう（設問 1）．次に，その回答にどの程度自信があるか回答してもらう（設問 2）．設問 2 は 1 4 の 4 段階（1:全く自信がない，2:あまり自信がない，3:少し自身がある，4:かなり自身がある）で回答してもらう．これを 86 対全てに対して行う．20 代男性 2 名が上記に従って 86 対を評価したものから，共に設問 1 で正解し，設問 2 で 4 と評価したものが使用する 26 対である．

5.2　変換手順

変換するのは F0 変化パターンである．以下の手順で変換を各文節で行う．

【step1】まず，入力音声を文節毎に分割する．i 番目の文節における F0 を $\{F0_i(t)\}$ とし，これを 3 次曲線に近似したものを $\{F0_i^{\mathrm{Ap}}(t)\}$ とする．オリジナルの F0 と近似した F0 の各点における差を $\mathrm{diff}_i(t) = F0_i(t) - F0_i^{\mathrm{Ap}}(t)$ とする．

【step2】$\{F0_i^{\mathrm{Ap}}(t)\}$ の平均値と極大値を第4章の分析結果に従い，$\{F0_i^{\mathrm{Ap}}(t)\}$ に α_L を加え，文節のラベルが Sc あるいは E ならば極大値をさらに 139.1 [cent] 高くする（図5.1）．$\{F0_i^{\mathrm{Ap}}(t)\}$ の最初のフレーム，最後のフレーム，$F0_i^{\mathrm{Ap}}(\tau)$，極小値の4点との2乗誤差平均が最小になるように3次曲線を生成する．

【step3】step2 で新たに得られた3次曲線に $\{\mathrm{diff}_i(t)\}$ を加えたものを $\{F0_i^{\mathrm{new}}(t)\}$ とする．従って，$\{F0_i(t)\}$ が持っていた微細変動は $\{F0_i^{\mathrm{new}}(t)\}$ でも保持し，大局的な変動のみが変換されたことになる．

【step4】入力音声における文節毎の音声と $\{F0_i(t)\}$ を使用してスペクトル包絡，非周期成分を求め，合成パラメータを定める．この合成パラメータのうち，F0 のみを $\{F0_i^{\mathrm{new}}(t)\}$ に差し替えて合成する．

5.3　変換精度と考察

客観評価と主観評価に分けて述べる，

5.3.1　F0 平均値の客観評価

演技音声（目標音声）の i 番目の文節における F0 を $\{F0'_i(t)\}$ とし，$\{F0_i(t)\}$，$\{F0_i^{\mathrm{new}}(t)\}$，$\{F0'_i(t)\}$ の時間方向の平均値を $\overline{F0_i}$，$\overline{F0_i^{\mathrm{new}}}$，$\overline{F0'_i}$ とする．

$\{F0_i^{\mathrm{new}}(t)\}$ は入力音声を目標音声の方向へ F0 を変換したものなので，$\overline{F0_i^{\mathrm{new}}}$ は $\overline{F0_i}$ よりも $\overline{F0_i^{\mathrm{new}}}$ の方に値が近く，$\overline{F0_i^{\mathrm{new}}}$ と $\overline{F0'_i}$ の値は近いということが望ましい．従って，本稿では次の2つの対立仮説を検定する．

対立仮説 1: $\overline{F0'_i} - \overline{F0_i^{\mathrm{new}}}$ と $\overline{F0_i^{\mathrm{new}}} - \overline{F0_i}$ の差が，0 よりも小さい平均，分散未知の正規分布に従う母集団に由来する．

対立仮説 2: $\overline{F0'_i} - \overline{F0_i^{\mathrm{new}}}$ が，平均 0 の正規分布に従う母集団に由来しない．

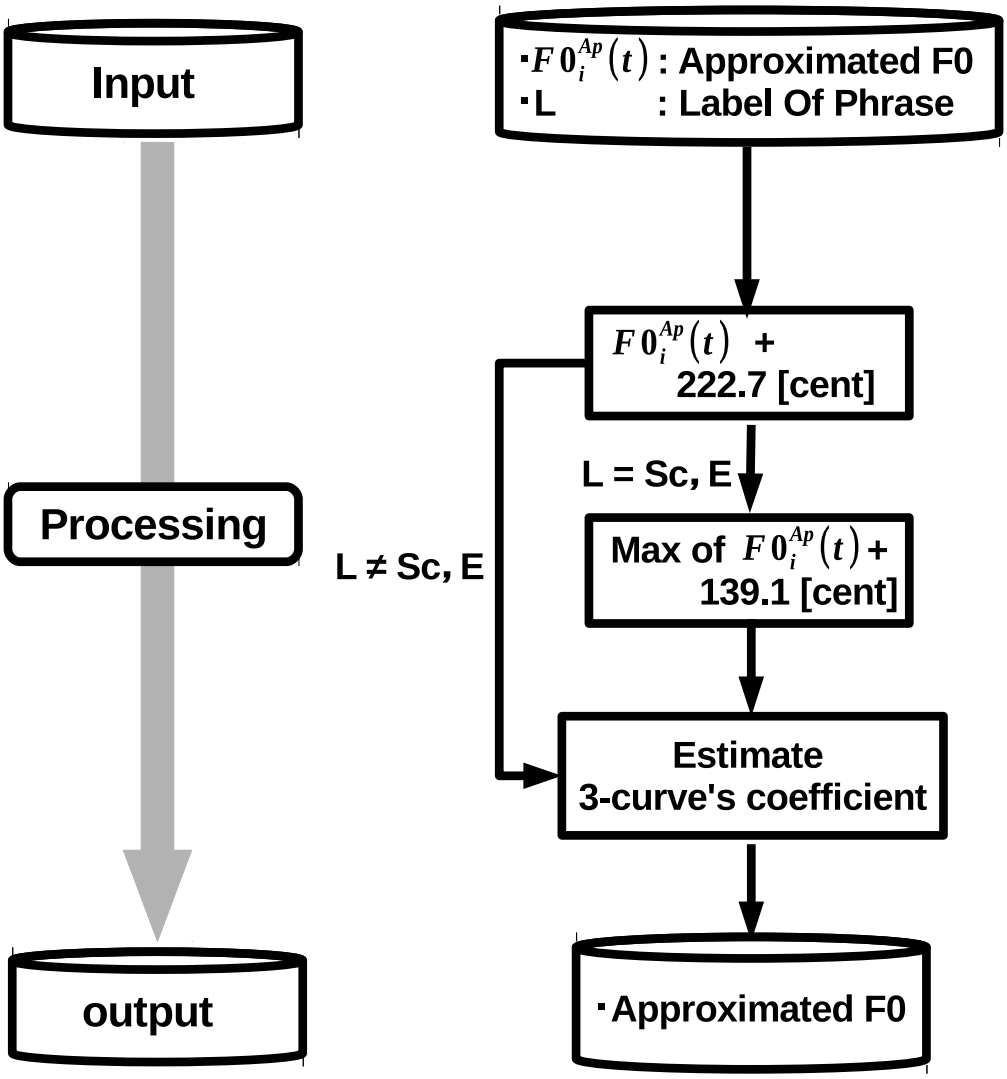
Input
Processing
output
・$F0_i^{Ap}(t)$: Approximated F0
・L : Label Of Phrase
$F0_i^{Ap}(t)$ + 222.7 [cent]
L = Sc, E
Max of $F0_i^{Ap}(t)$ + 139.1 [cent]
L ≠ Sc, E
Estimate 3-curve's coefficient
・Approximated F0

図 5.1: F0 変換の流れ

表 5.1: 平均値の差の検定結果

対立仮説 1			対立仮説 2
p 値	$D1$	$D2$	p 値
0.0000	27.5	25.8	0.0138

表5.1は上記の検定結果である．$D1$ は $\overline{F0_i^{\mathrm{new}}}$ と $\overline{F0'_i}$ の距離の平均値，$D2$ は $\overline{F0_i^{\mathrm{new}}}$ と $\overline{F0_i}$ の距離の平均値である．

対立仮説 1 が採択されたことで，$\overline{F0_i^{\mathrm{new}}}$ は $\overline{F0_i}$ より $\overline{F0'_i}$ に近いということが言える．また，対立仮説 2 が採択されたことで，$\overline{F0'_i}$ と $\overline{F0_i^{\mathrm{new}}}$ の差は平均 0 の正規分布の母集団に由来しているとは言えない．実際に，$\overline{F0_i^{\mathrm{new}}}$ と $\overline{F0'_i}$ の距離の平均値は 27.5 [cent] 程度である．これは，分析では 235 文節を使用しており，この数では十分ではなく変換時に $F0_i(t)$ に加える 222.7 [cent] が汎用性に欠けてしまったと予想できる．しかし，話者が見本音声なしに，地声と演技音声の音高を正確に，常に 222.7 [cent] の差が開くように話し続けることは困難であり，別日に音声を収録したことからも 27.5 [cent] 程度のズレが生じることは，本手法では防げないことである．

5.3.2　F0 極大値の客観評価

入力音声，変換音声，目標音声の i 番目の文節における F0 の極大値をそれぞれ $F0_i(\tau)$，$F0_i^{\mathrm{new}}(\tau^{\mathrm{new}})$，$F0'_i(\tau')$ とする．

$F0_i^{\mathrm{new}}(\tau^{\mathrm{new}})$ に関しても，5.3.1 節と同じくことが言えるので，本稿では次の対立仮説を検定する．

対立仮説 1: $F0'_i(\tau') - F0_i^{\mathrm{new}}(\tau^{\mathrm{new}})$ と $F0_i^{\mathrm{new}}(\tau^{\mathrm{new}}) - (F0_i(\tau) + 222.7)$ の差が，0 よりも小さい平均，分散未知の正規分布に従う母集団に由来する．

表 5.2: 極大値の差の検定

対立仮説 1			対立仮説 2
p 値	d1	d2	p 値
0.6106	32.4	7.0	0.2887

対立仮説 2: $F0'_i(\tau') - F0_i^{\mathrm{new}}(\tau^{\mathrm{new}})$ が，平均 0 の正規分布に従う母集団に由来しない．

表 5.2 は上記の検定結果である．$d1$ は $F0_i^{\mathrm{new}}(\tau)$ と $F0'_i(\tau)$ の距離の平均値，$d2$ は $F0_i^{\mathrm{new}}(\tau)$ と $F0_i(\tau)$ の距離の平均値である．対立仮説 1 は採択されなかったことで，$F0_i^{\mathrm{new}}(\tau^{\mathrm{new}})$ は $F0_i(\tau) + 222.7$ より有意に $F0'_i(\tau')$ に近いとは言えない．実際に距離の平均値を見てもわかる．また，対立仮説 2 に関しては，採択されなかったことで $F0_i^{\mathrm{new}}(\tau^{\mathrm{new}})$ と $F0'_i(\tau')$ の平均値の差に有意な差があるとは言えないので，これらは近い値を取るかもしれないということがわかる．

上記の対立仮説 1 に関して，より詳しく検定を行うため，次の対立仮説を検定する．

対立仮説 3: 先の対立仮説 1 において，ラベル Sc あるいは E のとき，それ以外のラベルのときに分け，ラベル群ごとの極大値で同じ対立仮説を立てる．

表 5.3 は対立仮説 3 の検定結果である．ラベル Sc，E にでは $d1$，$d2$ ともに表 5.2 で見たものよりも大きくなっているが，$p > 0.05$ であるため，対立仮説 1 と同じく，$F0_i^{\mathrm{new}}(\tau^{\mathrm{new}})$ は $F0_i(\tau)$ より有意に $F0'_i(\tau')$ に近いと言えない．また，その他のラベル群では $d1$ はほぼ変化がない（これは極大値を操作していないからである）．以上から，極大値に関してはラベル Sc と E のみに 139.1 [cent] を極大値に加えるだけでは足りないということがわかる．

表 5.3: ラベル毎による極大値の差の検定

対立仮説 3			
ラベル	p 値	d1	d2
ラベル Sc，E	0.1849	96.8	39.0
上記以外のラベル	0.8278	33.3	0.0

表 5.4: 設問 1

A に近い	少し A に近い	わからない	少し B に近い	B に近い
1	2	3	4	5

5.4　主観評価

10 名（男性 7 名，女性 3 名）の大学生に，聴取実験を行った．まず，被験者には音声 A と音声 B を聴かせ，次に音声 C を聴かせた．ことのとき，音声 A と音声 B の一方は入力音声，もう一方は目標音声，音声 C は変換音声である．このとき音声 C は音声 A と音声 B のどちらに近いかを答えさせた（設問 1，表 5.4）．また，音声 C をかわいく感じたかを答えさせた（設問 2，表 5.5）．これを 26 種類の文全てに行った．

設問 1 について，評価値を音声 A と音声 B ではなく，入力音声と目標音声のどちらに近いかという評価値に直し，表 5.6 にまとめた．また，設問 2 の評価値を表 5.8 にまとめた．

表 5.5: 設問 2

かわいくない	あまりかわいくない	少しかわいい	かわいい
1	2	3	4

表 5.6: 設問 1 の評価値・改

入力音声に近い	少し入力音声に近い	わからない	少し目標音声に近い	目標音声に近い
1	2	3	4	5

表 5.7: 設問 1 の評価平均値と標準偏差

セット	平均値	標準偏差	セット	平均値	標準偏差
1	2.6	1.43	14	2.1	1.30
2	3.2	1.40	15	4.2	0.98
3	2.7	1.35	16	2.8	1.17
4	2.9	1.58	17	1.6	0.49
5	1.6	0.92	18	2.8	0.98
6	1.1	0.30	19	2.2	1.25
7	1.9	1.14	20	1.4	0.66
8	1.6	1.02	21	3.5	1.43
9	2.0	0.77	22	1.7	0.90
10	1.6	0.92	23	1.8	1.40
11	2.5	1.02	24	1.6	0.49
12	1.8	1.17	25	2.5	1.50
13	1.3	0.46	26	1.5	0.50

表 5.8: 設問 2 の評価平均値と標準偏差

セット	平均値	標準偏差	セット	平均値	標準偏差
1	3.2	0.60	14	3.0	0.63
2	2.4	0.49	15	2.1	1.04
3	2.6	0.66	16	2.6	0.80
4	2.6	0.92	17	2.6	0.66
5	2.1	0.83	18	2.4	0.80
6	2.4	0.66	19	1.7	0.78
7	2.3	0.90	20	2.5	0.67
8	2.7	0.90	21	2.8	0.98
9	2.5	0.81	22	2.6	0.92
10	2.4	0.49	23	2.2	0.75
11	2.4	1.02	24	2.3	0.78
12	2.6	0.92	25	2.3	1.01
13	3.2	0.98	26	1.9	0.70

表 5.7 より，出力音声のほとんどが目標音声ではなく入力音声の方に近いという評価であった．その中でも 15 番目セットでは，音声 C が目標音声に近いあるいは少し近いと答えた人が多かった．また，2 番目と 21 番目のセットでも目標音声に近いあるいは少し近いと答えた人がある程度いた（図 5.2）．i 番目のセットで，目標音声と出力音声に対し，各音声で文節毎の F0 の時間方向の平均値を求める．両音声間の F0 平均値ベクトル（サイズは文節数に等しい）の残差平方和の平方根を文節数で割ったものを dist1 とする．同様に，出力音声と入力音声のペアに対しても残差平方和の平方根を文節数で割ったものを算出し，それを dist2 とする．

i 番目のセットで，目標音声と出力音声に対し，各音声で文節毎の F0 の極大値を求める．両音声間の F0 極大値ベクトル（サイズは文節数に等しい）の残差平方和の平方根を文節数で割ったものを Dist1 とする．同様に，出力音声と入力音声のペアで求めたものを Dist2 とする．ただし，入力音声の極大値は，抽出した極大値に平均値の差として定めた 222.7 [cent] を加えたものとする．一部のセットからそれぞれの距離を表 5.9 にまとめる．Dist1 ならば目標音声と出力音声で，共に極大値を持つ文節がない場合は空白にしてある．Dist2 ならば出力音声と入力音声において同様である．18，25 では 2 や 15，21 番目のセットのように，dist1 と Dist1 が dist2 と Dist2 よりそれぞれ小さかった．従ってこれらの音声は出力音声が目標音声に近づいている証拠である．しかし聴取評価では 18，25 番目のセットはあまり近づいているとは言えない結果になっている．この原因の一つはアクセントの位置が異なっているからと考えられる．本研究では，平均値，極大値の高さにしか着目せず，極大値の時間方向の位置までは考慮していない．また，もう一つの原因として，入力音声と目標音声との間で感情の強さの差が大きく感じられてしまうことだと考えられる．また，6 や 13，20 番目のセットにおいては平均値の差を埋める時点で，その増加分 222.7 [cent] では足りていないということがわかる．

設問 2 の評価に関して，設問 1 の評価結果と設問 2 の評価結果の間に相関がなかった（相関係数:-0.0056）．従って，本研究で用いた音声では，入力音声がかわ

表 5.9: 目標・出力と出力・入力音声のユークリッド距離

セット	設問 1 の評価平均	dist1	dist2	Dist1	Dist2
2	3.2	50.0	99.6	32.6	128.6
15	4.2	98.8	157.5	86.6	157.5
21	3.5	81.9	157.5	95.6	157.5
18	2.8	59.9	102.7	151.8	128.4
25	2.5	49.2	157.5	50.5	157.5
6	1.1	110.4	108.7	47.7	114.8
13	1.3	207.2	122.6	203.6	132.3
20	1.3	220.5	222.7		222.7
全セット平均	2.2	154.3	150.0	163.3	162.5

いらしいかどうか，目標音声がかわいらしいかどうかを一意に区別できるかは難しいということがわかる．

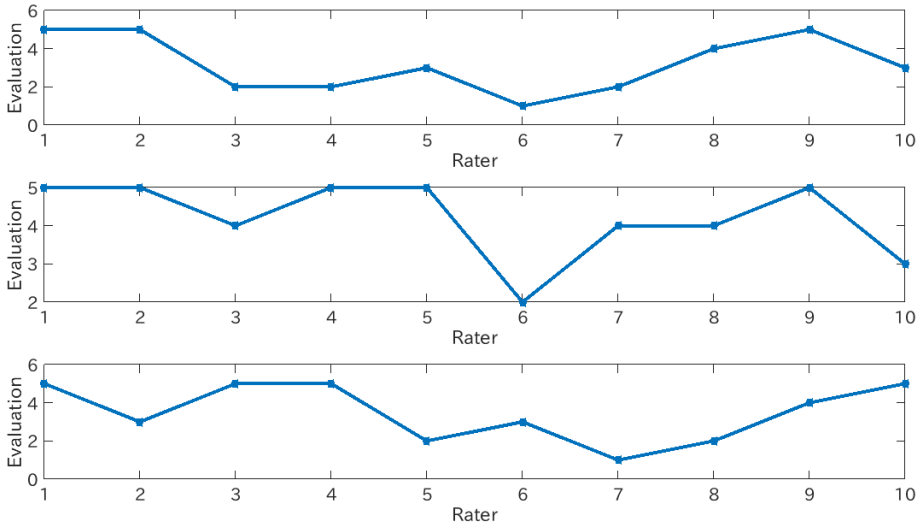

図 5.2: 上段：2 番目の音声　中段：15 番目の音声　下段：21 番目の音声

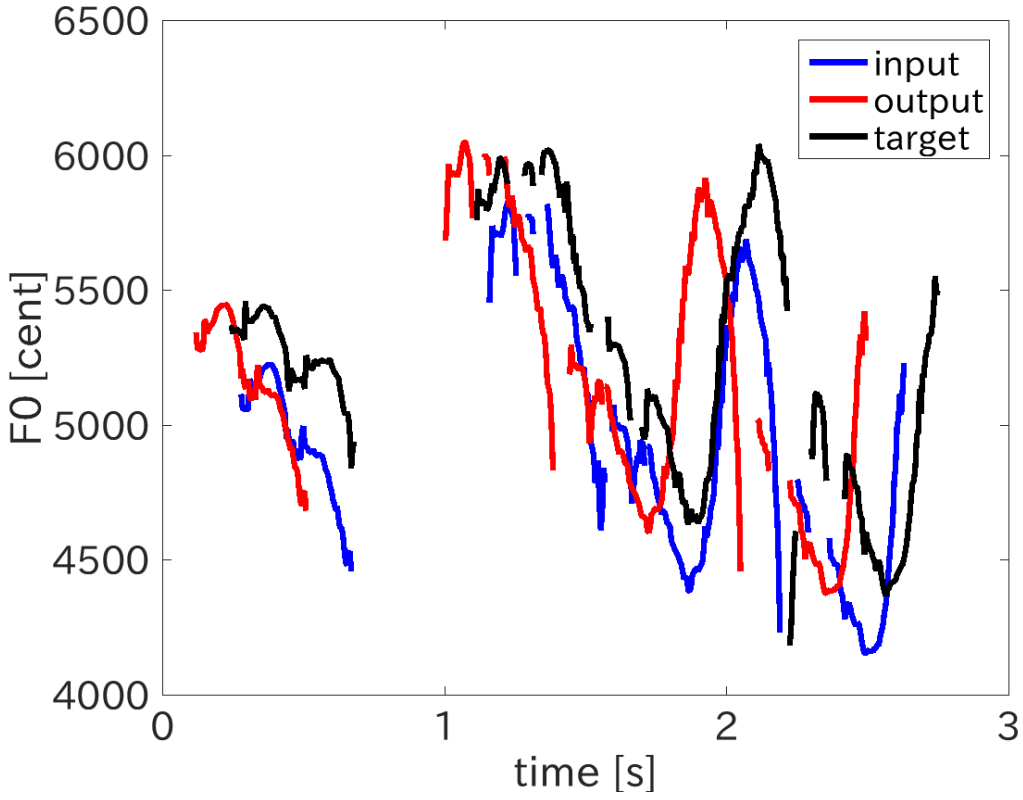

図 5.3: 3 番目のセットにおける 3 種類の音声の F0

(ねぇねぇ、さっきの 授業、何 やってたの？)

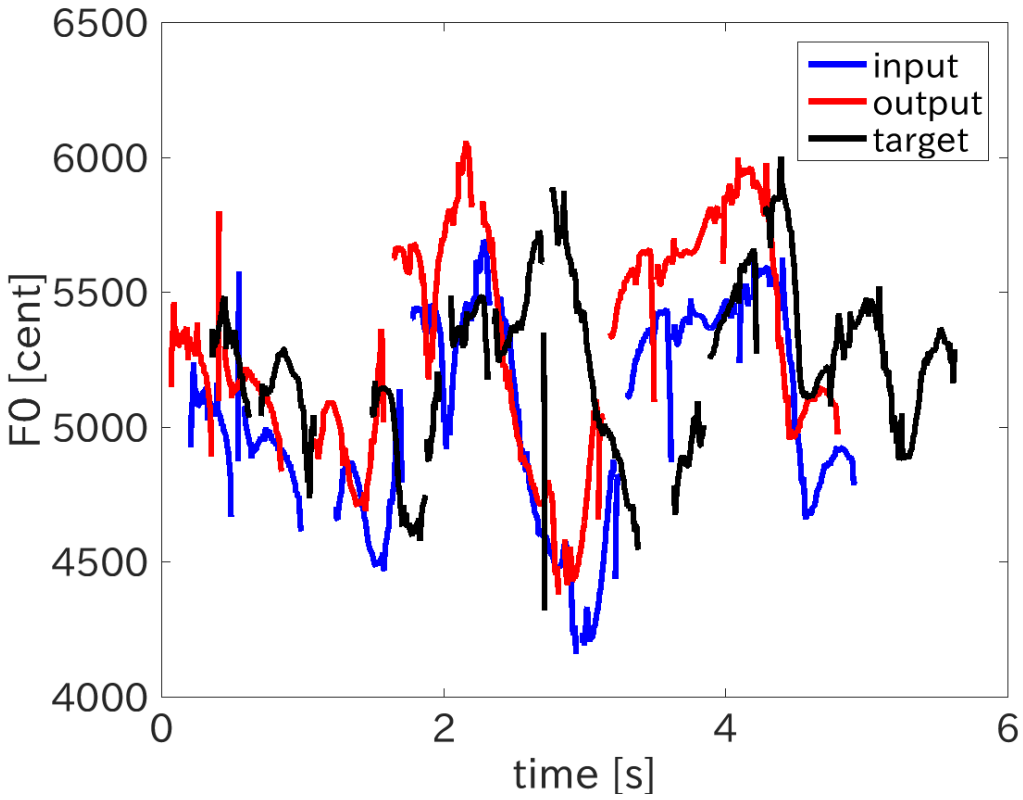

図 5.4: 18 番目のセットにおける 3 種類の音声の F0

(でもさぁ、もう 夜中の 12 時なのに、相変わらずの 夕日なんだね。)

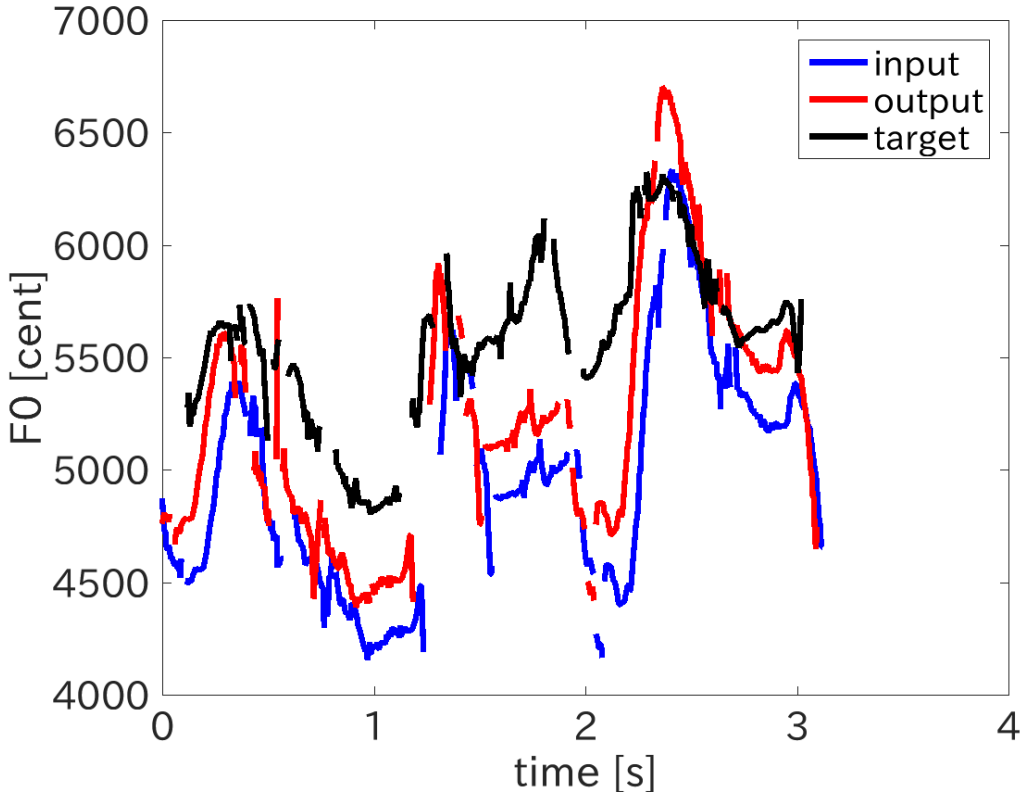

図 5.5: 13 番目のセットにおける 3 種類の音声の F0

(裕太くんったら 私が いないと ダメんだんだからぁ)

第6章　結論

本研究では，入力音声の韻律情報を操作することでかわいらしいと感じる音声に変換することを目指し，その第一段階として，地声と演技音声の間の F0 平均値と極大値を操作するだけでどの程度目的を実現できるかを調査した．文節を，その文節が属する文における位置によって分類し，その分類毎の文節で F0 の差を調べたところ，分類群に関わらず演技音声は地声音声よりもおおよそ 222.7 [cent] 高いということがわかった．また，その平均値の差を埋めても尚，極大値に有意な差がある文節も存在した．しかし，変換音声の聴取実験の結果，そのほとんどの音声の変換が不十分であるという結果になった．その原因として，分析に使用した音声データ数が十分でなかったこと，文章が感情を含んでいたにも関わらず分析や変換時に感情を考慮しなかったことが考えられる．また，分類した文節において，その前後の文節とのつながりを考慮しなかったことも原因と考えられる．

今後の課題としては，感情が含まれるような文章においては感情によっても分類を分けて分析する，前後の文節の種類による分析など，新たな考慮を入れて分析を行うということである．

参考文献

[1] 塩出 萌子，小泉 悠馬，伊藤 克且：“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（第 76 回全国大会），vol1，pp.495-496，2014.

[2] 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，pp.261-268，2005.

[3] 大谷 大和，戸田 智基，猿渡 洋，鹿野 清宏：“参照話者を用いた多対多固有声声質変換”，電子情報通信学会技術報告，SP2008-140，pp.85-90

[4] 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.

[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”，情報処理学会研究報告，2014-SLP-104(16)，1-6，2-14-12.

[6] 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”，電子通信学会論文誌，D-Ⅱ，Vol.J85-D-ll，No.4，pp.545-553，2002-4.

[7] Masato Akagi and Taro Ienaga：“Speaker individuality in fundamental frequency contours and its control”，J.Acoust．Soc．Jpn.(E)18，2，pp.73-80，1997.

[8] Elisabeth Zetterholm：“Same speaker - different voices Astudy of one impersonator and some of his different imitations”

[9] A.I.C.Monaghan，D.R.Ladd：“MANIPULATING SYNTHETIC INTONATION FOR SPEAKER CHARACTERISATION”

[10] 松原 実香，サトウ・タツヤ：“対象，評価，情動の観点から検討する「萌え」”，立命館人間科学研究，巻 26，pp.21-34，2013.

[11] 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 年春季研究講演論文集，No.2-Q5-22，pp.503-506，2014.

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

[13] 北村 達也 “物真似タレントによる物真似音声の分析”，信学技法，pp.49-54，2007.

[14] Cffon(レーベル): “お兄ちゃん CD”, 2006.

[15] 渋谷 貴紀，川端 豪：“音声対話システムのための親近感特徴量の探索”，電子情報通信学会技術報告，pp.25-30，2006.

[16] H.Fujisaki，“Vocal Physiology: Voice Production,Mechanisms and Functions”，Raven Press，pp.34-355，1988.

[17] 秋助：“夕映えワンダーホール”，“10 代から目指す！声優トレーニング最強 BIBLE”，pp.96-112，トランスワールドジャパン，2013.

[18] 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.

[19] 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・スペクトル包絡情報が品質の知覚に与える影響〜”，日本音響学会聴覚研究会，vol.41，no.7，pp.555-560，2011.

謝　　辞

本研究に取り組むにあたり，いつどんなときでも親身に指導してくださった北原鉄朗准教授には心より感謝しております．初めての研究ということもあり何もわからない私でしたが，いつも新しいことを教えていただき，その度に研究の楽しさ，この分野の学問の面白さを知ることができました．本当にありがとうございました．

山梨大学の森勢将雅特任教授には短い間でしたが，技術的なアドバイスをいただきました．普段得られない技術など様々なアドバイスや貴重な意見をくださり深く感謝致します。

ご自身の学業などで忙しい中，聴取実験などに協力してくださった皆様に深く感謝致します．特に，田島侑佳さんと中條沙織さんは音声データの録音にご協力してくださり，本当にありがとうございました．私の研究にはなくてはならない貴重なデータでした．

研究を始めた当初，いつでも一番近くでサポートしてくださった大塚匡紀さんに深く感謝致します．

最後に，居心地の良い環境にしてくださった同研究室の皆様と，学業に集中させてくださった両親に深く感謝致します．

本論文を作成できたのは皆様のおかげであり，皆様のご協力があってのことだと思っております．本当にありがとうございました．