平成 29 年度 修士論文

音声における「かわいさ」の制御のための声質変換

指導教員 北原鉄朗准教授

日本大学大学院総合基礎科学研究科

6116M07 大野涼平

2018 年 2 月 提出

概　　要

かわいい音声の生成は，動画制作などの二次制作等を始めとする，エンターテインメントの側面で貢献できると期待される．一方で，ある音声を聴取したとき，それをかわいいと判断するかどうかは聴取者に強く依存してしまう．従って，かわいい音声を生成するためにはユーザの思うかわいさをモデル化する必要がある．それを実現する上で通常の声質変換技術（発話内容を保持したまま入力音声の声質を目標話者のものに変換して出力する技術）の問題点は，目標音声の話者性とかわいさが分離されていないことである．つまり，ある話者らしさを残してかわいさだけを制御することができないことである．かわいさに影響する特徴のみを制御する手法が望まれる．これを実現するには，かわいさに影響する特徴のみを話者性とは別に制御する必要がある．そこで，本研究では，声質変換によるユーザの思うかわいい音声の生成を実現するため，ユーザの思うかわいさを学習し，かわいさに起因する特徴のみを変換可能なモデルを提案する．本研究における提案手法では，ユーザには予め多数の話者をかわいい話者とそうでない話者の 2 つのクラスに分類してもらい，それらの話者の音声データを用いてユーザの思うかわいさを分析する．

本研究ではまず，3 章で音声におけるかわいさの評価と聴取時間の関係を分析した．女性声優による女性声から 75 ms, 150ms, 300ms, 600ms の区間を切り出した音声とフルの音声に対してどれくらいかわいさを感じたか 5 段階評価の主観評価実験を行った結果，(1) 評価の安定性は聴取時間が短くなると低くなる，(2) 4 以上の評価を得たフル音声から切り出した音声に対して，聴取時間が短くなると評価の曖昧性は高くなる，(3) 2 以下の評価を得たフル音声から切り出した音声に対して，聴取時間が短くても評価の曖昧性は下がらない，(4) 75 ms でも，半数の音声については，評価の安定性はフル音声と同程度である，という結果が得られた．これは，特徴の大局的な時間変化をモデル化しなくても, 一定程度のかわいさの制

御は可能であることを示す.

4 章では，他ユーザのどの話者をどれくらいかわいいと感じるかという情報（かわいさの評価の傾向）を参照し，協調フィルタリングを用いて，ユーザにかわいいと思う話者の推薦を行い，その精度を調査した．被験者らには 70 名の女性声優の音声に対してかわいさの 7 段階評価を行わせ，その内 60 名分に対する評価を未評価とし，その評価値を予測，および，かわいい話者を推薦話者として出力した．その結果，(1) 評価値の予測精度である平均絶対誤差は 0.9 程度で，話者推薦における適合率の中央値は 0.8 であった，(2) 高評価を与えた話者の数と類似評価者数に正の相関があった (相関係数:0.649)，(3) かわいさの評価の傾向において，男女間で大きな差は見られなかった，という結果が得られた．

5 章では，ユーザの思うかわいい声の生成を実現するため，話者適応型 Restricted Boltzmann Machines を拡張し，かわいさの特徴のみを制御可能な声質変換モデルを提案する．ユーザによって分類されたかわいい話者とかわいくない話者の 2 つのクラスに依存する特徴，つまり，かわいさの特徴を表現するパラメータを定義する．このパラメータを適切に変えることで，かわいさの特徴のみの制御を実現できる．従来の声質変換によって変換した音声と提案法でかわいさのみを制御した音声のかわいさを順位付けさせ，また，提案法の音声のかわいさと話者性について 7 段階の主観評価を行った結果，提案法では (1) 従来の声質変換手法よりもかわいさが優位で，(2) 60% 以上の評価が変換前よりもかわいくなり，(3) 変換音声の話者性が 44% の評価でかわいいと思う話者の話者性に近い，という結果が得られた．

目　　次

図 目 次

表 目 次

第1章　序　　論

1.1　はじめに

女性声における「かわいさ」は日本のポップカルチャーで注目されている．例えば，日本のアニメーションにおいては，女性声優によるかわいらしい声，あるいは女性らしさ・女子らしさを強調したような声，媚びたような声が人気を高めている要因の一つとして挙げられる．メイドカフェ（コスプレレストラン）においては，女性メイドがかわいらしさを強調した喋り方を使用している．これらの声は“萌え声（萌え：強い愛着を表す，インターネットスラング）”と呼ばれることも多い．一方で，音声合成技術や声質変換技術の発展に伴い，合成音声が動画制作などの二次制作に応用されており，エンターテインメントの側面でユーザの自由度は高まってきている．そうした中,「かわいい声」を合成することもまた，エンターテインメントの側面で有用だと考えられる．

かわいい声を生成するには，音声におけるかわいさの程度の評価と音響特徴量の関係性の分析が必要である．しかし，かわいさを感じるかどうか（例えば，ある“モノ”に対して，それをかわいいと思うかどうか）は個人の価値観によって異なり，同じ音声に対してもかわいいと思うかどうかは聴取者に依存する．そのため，萌え声やかわいい声とはどのような声かと言った議論は容易ではない．そこで，かわいい声を生成するためには，ユーザがどの音声をどれくらいかわいいと感じるかという情報（かわいさの評価の傾向）に沿ったモデルの構築が必要である．例えば，予めユーザには多数の話者を，かわいい声の話者とそうでない話者に分けてもらい，その 2 種類の話者間の違いをモデル化することで，そのユーザが思うかわいさを分析できる．一方で，ユーザ間でかわいさの評価の傾向に類似性が存在すれば，他ユーザの評価の傾向から，当該ユーザが思うかわいい話者を推定することができる．そうすれば，ユーザは少数の話者を分類するだけで，多

数の話者の分類を行うことが可能である.

1.2 本論文の目的

本論文では，声質変換によるユーザの思うかわいい声の生成の実現を目的とする.

まず，聴取した音声がかわいいかどうかという判断は，音声の局所的な特徴やスペクトルが参照できれば良いのか，あるいは特徴の大局的な時間変化や韻律を参照する必要があるのかを調査し，局所的特徴を参照することでかわいさの判断ができることを示す.

次に，ユーザ毎にかわいさの感じ方が異なることへ対応するため，ユーザには予め多数の話者をかわいい話者とそうでない話者に分けさせ，その 2 種類の話者でモデルを構築する．その際に，かわいさの評価の傾向を利用して少数の話者を分類するだけで多数の話者の分類を可能とすることを目指す．そこで，評価者間でかわいさの評価の傾向に類似性があると仮定し，協調フィルタリングを用いた評価値の予測・かわいい話者の推薦を試みる.

声質変換については，話者適応型 Restricted Boltzmann Machines[21] を拡張し，かわいさの制御を行うモデルを提案する．実験的評価では，提案モデルによってかわいさの特徴を表現するパラメータの推定が可能であることを示す.

1.3 本論文の構成

本論文の構成は次の通りである．第 2 章では，女性声の魅力やかわいさに関する従来の調査について，および，従来の統計的声質変換技術について述べ，本研究で取り組む課題についてを述べる．第 3 章では聴取時間の長さによる音声におけるかわいさの近くに与える影響について調査し，その結果を述べる．第 4 章では評価者間でのかわいさの評価の傾向，および，その類似性について調査し，その結果について述べる．そして，第 5 章では話者適応型 Restricted Boltzmann Machines[21] を拡張した，かわいさの制御を行うモデルを提案する．最後に，第 6 章で本論文のまとめを記す.

第2章　従来研究と課題

女性声におけるかわいさ，および，かわいさに関連する印象に対する主観評価と音響特徴量の関係性における調査について述べる．次に，従来の声質変換技術について述べ，最後に本研究において取り組むべき課題について述べる．

2.1　女性声の魅力と音響特徴量に関する調査

女性声の魅力の知覚と音響特徴量との関係性について，多くの調査がされてきた [1]．Jones らは，女性声は，基本周波数（F0）が高い方が男性に好まれると述べた [2]．Lie らは，F0 だけでなく，音声の息づかいの重要性を述べた [3, 4]．また，単母音と単語では，音声における魅力の程度が異なるという報告もある [5]．Babel らは女性声への魅力の評価において，男女の評価者間で正の相関があると述べた．また，彼らは，女性声の魅力を測定するための音響特徴量として声道長に関連する特徴量を用いることが有用だと報告している [6]．Puts らは，フォルマント周波数のバラつきが女性声の魅力に貢献すると述べた [7]．また，女性はそのフォルマント周波数のバラつきや変化に気付きやすいとも述べており，女性は男性が好む声の特徴を再現するように使用している事を示唆した．高野らは “萌え声” に着目し，ある音声から萌えを知覚する（萌える）かどうかは，その音声における感情表現など韻律よりも，個人性に依存すると述べた．また，萌える音声においては，韻律（例えば F0 の平均値や分散値，話速度など）を制御する事によって，萌えの程度が変化するということを述べている [8, 9]．川原によるメイドカフェ店員の演技声と地声の違いに関する調査では，演技声は地声よりも F0 が高く，話速度も高かった [10, 11]．Starr は “Sweet Voice” について，その知覚には声のクリーキーさや息づかいと関係があると述べた [12]．

2.2　統計的声質変換

統計的声質変換とは，発話内容を保持したまま，入力音声の声質を目標話者のものへと変換する技術である．変換元話者と目標話者の発話データを用いて，両話者の対応関係を学習するために様々な手法が提案されてきた．その中の有名な手法の 1 つに Gaussian Mixture Model (GMM) に基づくモデル（GMM 変換法）がある [13, 14, 15]．一般的な GMM 変換法は，元話者と目標話者の同一発話内容の音声データから得た音響特徴量を時間同期した音響特徴量（パラレルデータ）を用いて GMM により結合確率密度関数をモデル化する．元話者と目標話者の音響特徴量の静的・動的特徴量ベクトル $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ ($\boldsymbol{x}_t,\ \boldsymbol{y}_t$ はフレーム t における音響特徴量で，$\Delta\boldsymbol{x}_t,\ \Delta y_t$ はそれぞれの動的特徴量) を結合したベクトル $\boldsymbol{Z}_t$ を GMM でモデル化する．

$$p(\boldsymbol{Z}_t|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{Z}_t;\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.1}$$

このパラメータ $\boldsymbol{\theta}$ は EM アルゴリズムによって推定される（学習）．変換時は，元話者の音声の音響特徴量からを最尤系列推定法 [14] によって目標話者の音声の音響特徴量を推定する．元話者の音声の静的特徴量系列 $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top$ と学習済みのパラメータ $\boldsymbol{\theta}$ を用いて，出力音声の静的特徴量系列 $\hat{\boldsymbol{y}}$ を以下の式で推定する．

$$\begin{aligned} \hat{\boldsymbol{y}} &= \underset{\boldsymbol{y}}{\operatorname{argmax}}\, p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X}) \\ &= \underset{\boldsymbol{y}}{\operatorname{argmax}}\, p(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{W}\boldsymbol{x}) \end{aligned} \tag{2.2}$$

ただし，$\boldsymbol{W}$ は $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{x},\ \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ を満たす静的・動的制約行列である．

また，近年は Deep Neural Networks (DNN) に基づくモデル（DNN 変換法）が GMM 変換法よりも精度が高いことで注目を浴びている [16, 17]．こちらは Feed Forward 型ネットワークによる推論で元話者の静的・動的特徴量系列 $\boldsymbol{X}(= \boldsymbol{W}\boldsymbol{x})$ から得られる推定した目標話者の静的・動的特徴量系列 $\hat{\boldsymbol{Y}}$ に対して，最尤パラメータ生成 [18] によって求まる静的特徴量系列 $\hat{\boldsymbol{y}}$ と目標話者の静的特徴量系列 $\boldsymbol{y}$ の誤差を最小化するようなネットワークを学習する．

GMM 変換法も DNN 変換法も，どちらも元話者と目標話者のパラレルデータを必要とするため，学習データを準備するためのコストが高いことが問題視されている．そこで，パラレルデータを用いない手法として固有声変換法 [19] や，Variational Auto-Encoder (VAE 変換法)[20] や Restricted Boltzmann Machines （話者適応型 RBM 変換法 [21]）を用いた手法が提案されており，変換対象の話者について柔軟性が高まってきている．

2.3 課題と解決策

声質変換では，主に韻律などの大局的特徴の変換とスペクトルなどの局所的特徴の変換から成る．特に，スペクトル変換による声質変換の多くは短時間フレーム毎に元話者と目標話者間の対応関係を学習している．声質変換によって個々のユーザが思うかわいい声を生成するためには，第一に，大局的な特徴（広い時間幅の時間変動）と，スペクトルなど局所的な特徴のどちらの影響がよりかわいさを感じるために重要か議論する必要がある．もし，局所的な特徴を参照するだけでかわいさの評価が可能であれば，従来のスペクトル変換による声質変換と同様の枠組みでかわいい音声への変換が可能である．

また，先に述べたように，ユーザ毎にどのような音声をかわいいと感じるかどうかは意見が異なるため，ユーザのかわいさの評価の傾向に依存したモデルを構築する必要がある．そのための方策の一つとして，ユーザに予め多数の話者から “かわいい話者” と “かわいくない話者” の 2 クラスに分類させ，クラス内で共通な特徴やクラス間で異なる特徴を分析することが挙げられる．そのとき，もしもユーザ間でかわいさの評価の傾向に類似性があるならば，他ユーザの評価の傾向を用いることで，多数の話者ではなく少数の話者を分類するだけで多数の話者の分類を可能とすることが期待できる．

従来の声質変換技術では，ユーザがかわいいと思う話者を目標話者に設定することで，その話者の声質をもつ音声を生成することができ，かわいい音声を得ることができる．しかし，変換元話者に様々な話者を想定しても，出力される声質は目標話者ただ一人のものに限定されてしまう．本来は，入力音声をかわいく加工した音声が出力されることが望まれ，声質そのものを全てある一人の人物のも

のとして出力してしまうのは応用の幅を限定してしまう.

以上から，声質変換によって個々のユーザが思うかわいい声を生成するための課題を 3 つに要約できる.

(1) かわいさの評価と聴取時間の関係性
(2) ユーザが思うかわいい話者の推薦
(3) かわいさのみの制御

本研究では課題 (1) について，音声におけるかわいさの評価が聴取時間の変化にどのような影響を受けるかを，聴取実験によって分析する．課題 (2) については，評価者間でかわいさの評価の傾向に類似性があると仮定し，協調フィルタリングを用いた評価値の予測・かわいい話者の推薦を試みる．最後に課題 (3) については，話者適応型 Restricted Boltzmann Machines[21] を拡張したモデルを提案する．提案モデルは，話者に起因する特徴の内，かわいさの知覚に強く影響する特徴とそうでない特徴を区別するようなモデルであり，強く影響する特徴のみを制御することで，かわいさのみの制御の実現が期待できる．尚，以降では，スペクトル変換による声質変換を単に声質変換と呼ぶ.

第3章　女性声のかわいさの評価と時間長の関係性

本章では，音声のかわいさのモデル化に先立ち，音声に対してかわいさを評価するための音響的条件について検討する．特に，音声の聴取時間に着目し，極めて短い音声しか聴取しない場合にかわいさを感じることができるか調査する．

3.1　はじめに

ある音声をかわいく感じる場合，何の特徴量に対してかわいさを感じたのかについては，様々な可能性が考えられる．まず，基本周波数（F0）の平均が高いときや，その分散が大きい方が女性声の魅力が高くなることが知られている [2, 8]．その他にも，声道長は幼さに対応するため，これもかわいさのモデル化に有効な可能性がある [6]．また，これらの時間変化についてもかわいさに関係ある可能性がある．いわゆるメイド声は地声よりもアクセントにおける F0 をより上げる傾向が見られ [10]，F0 の時間変化がかわいさのモデル化に有効な可能性がある．

このような様々な可能性がある中で，ここでは，瞬間的な特徴だけでかわいさの評価が可能なのか，ある程度長い音声を聴かないとかわいさの評価が可能ではないのかを調査する．具体的には，女性声の音声信号から 75~600ms の区間を切り出し，被験者に聴いてもらってそのかわいさを評価してもらう．その評価の安定性（同じ音声に対する評価が変動する度合い）と曖昧性（音声に対する評価が「どちらとも言えない」に近づく度合い）が，切り出す前の音声と比べて高くならないのであれば，かわいさは音声の局所的な特徴に起因すると考えられる．一方，切り出す前の音声より高くなるのであれば，一定の長さを聴かないとかわいさの評価ができないことを意味し，スペクトルや F0 の時間的な変化がかわいさの評価

表 3.1: 音声刺激の内訳

Duration	Cut-out position	For all actress
75 [ms]	3 patterns (random)	36
150 [ms]	3 patterns (random)	36
300 [ms]	3 patterns (random)	36
600 [ms]	3 patterns (random)	36
Full [ms]	1 pattern (no cut-out)	12
Sum	-	156

に重要であると示唆される．その場合，音声のかわいさを制御するには，フレーム単位の特徴だけでなく，大局的なモデル化が必要であることになる．

3.2　実験・条件

3.2.1　実験データ

本実験では “お兄ちゃん CD” [22] に収録されている音声を使用する．当 CD には 12 名の女性声優それぞれが 100 個のシチュエーションで「お兄ちゃん」と呼びかける音声が収録されている．この内，高野らの萌え声に着目した実験 [8] で萌えが強かった “ちょっと楽しい” というシチュエーションの音声を使用する．

12 個の音声それぞれから，75 ms，150 ms，300 ms，600 ms だけ 3 つずつ音声を切り出し，それらの切り出し音声と，切り出しを行わない 1 秒程度の源音声（フル音声）を音声刺激とする．ただし，切り出し位置はランダムとし，切り出し音声に無音区間が含まれた場合は位置を変えて選びなおす．従って，音声刺激の内訳は表 3.1 の通りになる．

3.2.2 実験手順

14 人の日本人男女（男女 7 名ずつ，21 歳 ~24 歳，平均:21.43 歳，標準偏差:0.94 歳）に，以下の通りの実験を行った．

1. 75-ms の音声刺激について，
1-1. 36 種（切り出し位置 3 パターン × 声優 12 名分）を 3 つずつ計 108 個の音声刺激をランダム順に聴取する．同一の刺激を 3 つ使用するのは，同一の刺激に対する評価の安定性を測るためである．
1-2. 各刺激を聴取した後，そのかわいさを 5 秒以内に 5 段階で評価する（1:かわいくない，5:かわいい）．
1-3. 54 個を評価した時点で休憩をする．
2. 150-ms, 300-ms, 600-ms においても同様に行う．
3. フル音声 36 個に対しても同様に行う．

3.2.3 分析方法

本実験では，音声の聴取時間とかわいさの評価の安定性及び曖昧性の関係を調査する．このとき，安定性は評価のバラつきによって表現され，曖昧性は “3” と評価された割合で表現されると考えられる．従って，以下の 2 つの方法で分析を行う．

安定性に関する分析 同一の音声刺激に対して評価者は 3 回評価を行っている．この 3 回の評価値のレンジ（最大値と最小値の差）が大きければその評価者の当該刺激に対する評価の安定性は低いと見なせる．従って，音声刺激の時間長と評価値レンジの関係性を分析する．

曖昧性に関する分析 フル音声に対して 3 回とも 4 以上の評価を与えたにも関わらず，その切り出し音声に対する 3 回の評価が “3” に近ければ，あるいは低評価になれば，その切り出し音声に対する評価の曖昧性が高いと見なせる．従って，音声刺激の時間長と切り出し音声に対する評価値の中央値（ただし，当該切り出し音声のフル音声に対して 3 回とも 4 以上を与えたものに限る）の関係性を分析する．また，同様にフル音声に対して 3 回とも 2 以下を与えた場合に対しても分析を行う．

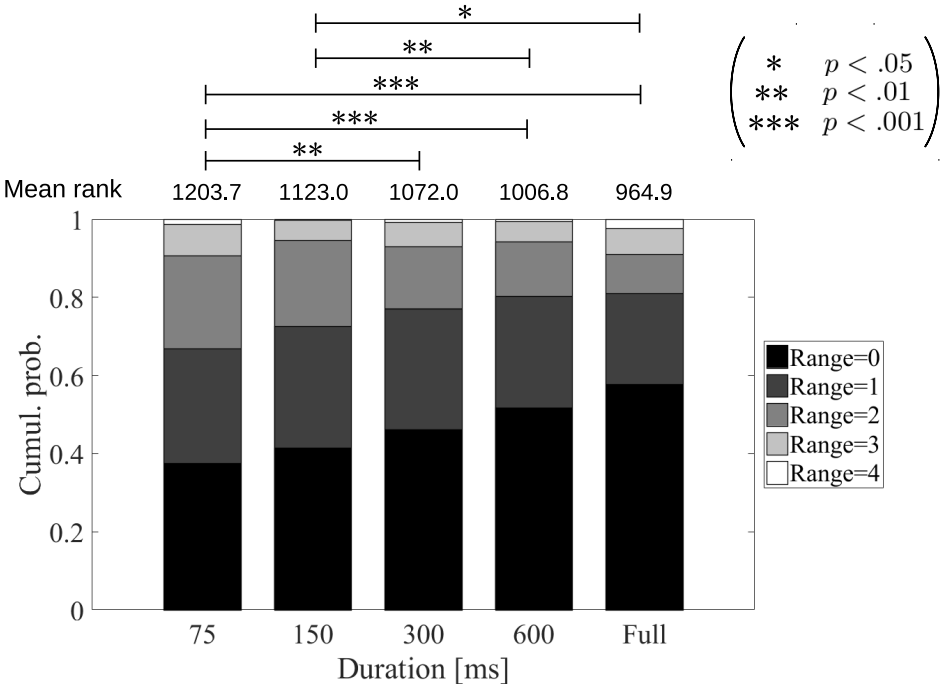

図 3.1: 同一刺激に対する評価におけるレンジの累積度数分布

3.3　実験結果・考察

3.3.1　安定性に関する分析の結果

同一刺激に対する 3 回の評価のレンジの分布を図 3.1 に示す．図 3.1 からは次のことがわかる．

– 同一刺激に対して 3 回とも同じ評価値が与えられた割合は全体で 46.9% である．
– 同一刺激に対する 3 回の評価のバラつき（レンジ）は，音声刺激の時間長が短くなるにつれ大きくなっている．
– 75 ms でも，66.9% の音声刺激に対してレンジが 1 以下であった．

Kruskal-Wallis H test (KW-test, $\alpha = .05$) によって，5 つの分布間で平均順位に有意な差がないという帰無仮説を検定する．各分布のサンプルサイズは 504 個で，フル音声における分布は 168 個である．その結果（$H(df = 4) = 39.04, p < .001$），5 つの分布の内，少なくとも 2 組間には有意な差があるということがわかった．それを踏まえ，次に，Mann-Whitney U test (U-test, $\alpha = .05$) による多重比較

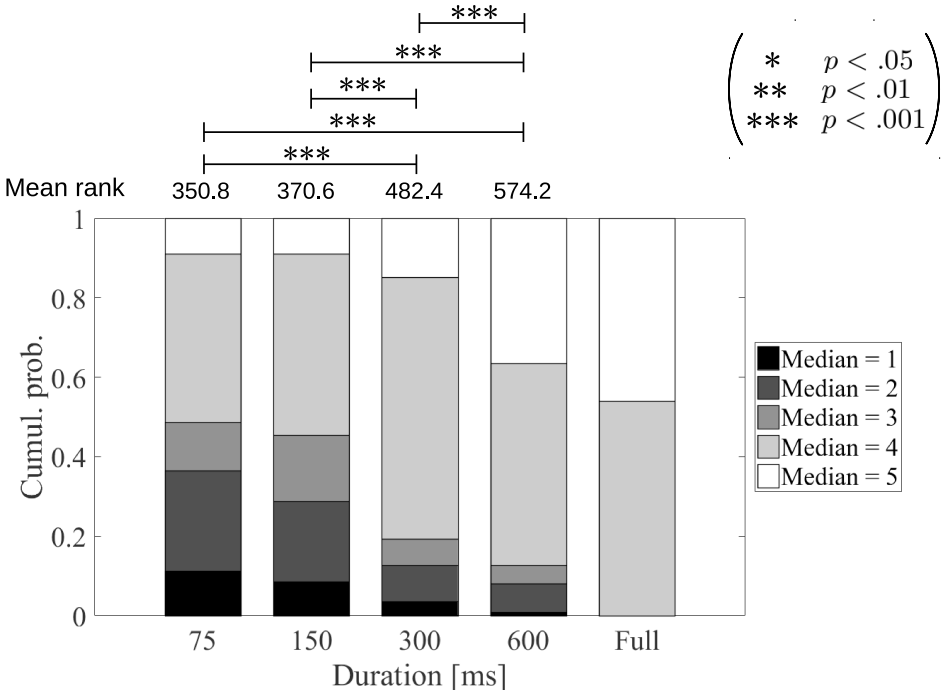

図 3.2: 高評価のフル音声の切り出し音声に対する評価における中央値の累積度数分布

(Bonferroni 法による補正有り) で 2 組間の分布の差を検定した．その結果，75 ms と 300 ms ($z = 3.53, p < .01$)，75 ms と 600 ms ($z = 5.25, p < .001$)，75 ms とフル音声 ($z = 4.33, p < .001$)，150 ms と 600 ms ($z = 3.39, p < .01$)，150 ms とフル音声 ($z = 3.15, p < .05$) で有意な差が見られた．従って，かわいさの評価の安定性は聴取時間が短くなるにつれて下がっていくが，一方で，聴取時間が短くても安定性が低いとは言いがたい．

3.3.2 曖昧性に関する分析の結果

フル音声に 4 以上の評価を与えた音声に対する分析

本分析では，同一のフル音声に対して 4 以上の評価を与えた場合の，その切り出し音声に対する評価値の中央値を分析する．その中央値の分布を図 3.2 に示す．図 3.2 からは次のことがわかる．

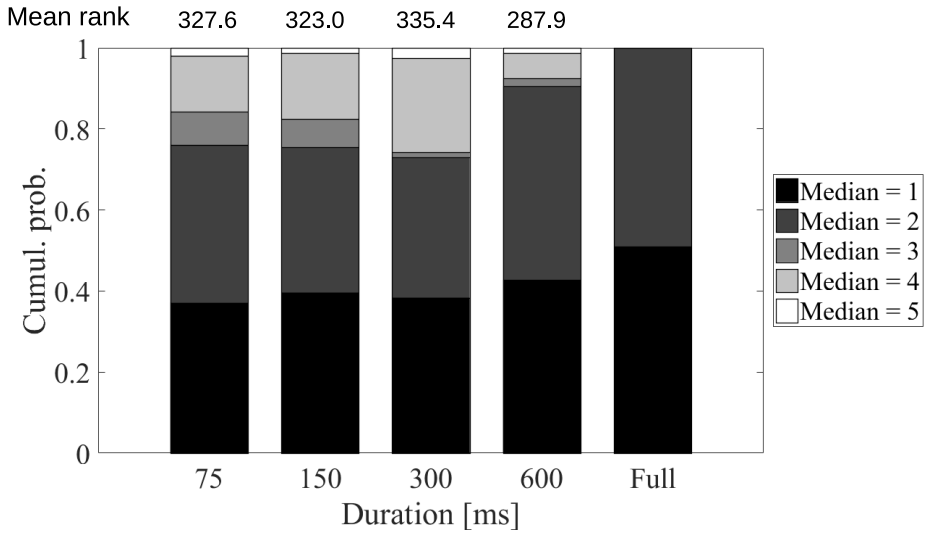

図 3.3: 低評価のフル音声の切り出し音声に対する評価における中央値の累積度数分布

– フル音声に対して 4 以上の評価を与えているにも関わらず，時間長が短くなると低評価を与える割合が増えている．
– 75 ms でも，51.4% の音声刺激に対する評価がフル音声に対する評価と同様に高い評価値を与えている．

フル時間における分布は，評価値が 3 回とも 4 以上である場合のみを対象としているため，フル時間における分布を除いた 4 つの分布間で検定を行う．各分布のサンプルは 222 である．KW-test の結果（$H(df = 3) = 128.01, p < .001, \alpha = .05$）に基づき，前節同様に多重比較を行った．その結果，75 ms と 300 ms ($z = -6.12, p < .001$)，75 ms と 600 ms ($z = -9.34, p < .001$)，150 ms と 300 ms ($z = -5.40, p < .001$)，150 ms と 600 ms ($z = -8.84, p < .001$)，300 ms と 600 ms ($z = -4.81, p < 001$) で有意な差があるとわかった．

フル音声に 2 以下の評価を与えた音声に対する分析

本分析では，同一のフル音声に対して 2 以下の評価を与えた場合の，その切り出し音声に対する評価値の中央値を分析する．その中央値の分布を図 3.2 に示す．図 3.3 からは次のことがわかる．

– 図 3.3 とは異なり，時間長と評価値の分布に相関は見られなかった．
– 75 ms でも，76.1% の音声刺激に対する評価がフル音声に対する評価と同様に低い評価値を与えている．

前節同様に，フル音声における分布を除いた 4 つの分布間で検定を行う．各分布のサンプルは 159 である．KW-test（$\alpha = .05$）の結果，$H(df = 3) = 7.16, p = .067$ であり，有意な差は見られなかった．

3.3.3 考察

局所的特徴と大局的特徴

聴取時間が短くなれば評価の安定性が低くなるという本実験の結果は，女性声の魅力や萌えに対する主観評価は，韻律など音響特徴量の時間変動（F0 軌跡や分散など [10, 11]）の影響を強く受けるという報告と一致する．一方で，4 以上の評価を与えたフル音声の切り出し音声に関して，51.4% の評価もまた 4 以上であった．大局的特徴を操作せずとも，局所的特徴だけでもかわいさを制御できると言える．

「かわいい」と「かわいくない」の非対称性

フル音声に 4 以上の評価をしていても，時間長が短くなるとその音声に対する評価の曖昧性が高くなる．一方で，このような傾向は，フル音声に 2 以下の評価をしていた場合では見られず，75 ms という短い音声に対しても 76.1% が 2 以下の評価であった．この違いは評価の非対称性が原因と考えられる．一般的に，主観評価では対称な用語を使用する（例：“明るい”，“暗い”）が，“かわいい” と対称な用語が存在しないため，今回は “かわいくない” というような否定語を使用している．従って，評価者は音声刺激に対してネガティブな要素を感じた場合に限らず，かわいさを感じることができなかった場合にも “1” や “2” を付ける事が考えられる．この非対称性が上記の分析結果の違いの要因の 1 つであると考えられる．

3.4　本章のまとめ

本章では，音声におけるかわいさの評価と，聴取時間の長さとの関係性を調査した．その結果，

(1) 評価の安定性は聴取時間が短くなると低くなる．
(2) 4 以上の評価を得たフル音声から切り出した音声に対して，聴取時間が短くなると評価の曖昧性は高くなる．
(3) 2 以下の評価を得たフル音声から切り出した音声に対して，聴取時間が短くても評価の曖昧性は下がらなかった．
(4) 75 ms でも，評価の安定性（あるいは曖昧性）は低い（高い）とは言い難い．

よって，特徴の大局的な時間変化をモデル化しなくても，一定程度のかわいさの制御は可能であると期待できる．

第4章　音声における「かわいさ」の主観的評価値の自動推定

本章では，かわいさについて，どの音声にどれくらいの評価を付けるか（かわいさの評価の傾向）を協調フィルタリング [26] により推定する方法について述べる．

4.1　はじめに

個々のユーザにとってかわいい音声を生成するには，そのユーザがどの音声にどの程度のかわいさを感じるか（かわいさの評価の傾向）の情報が必要である．これは，かわいさの評価の傾向は，人によって異なると予想されるからである．しかし，個々のユーザが十分な個数の音声に対してかわいさの評価を行うのは容易ではない．音声には一覧性がなく，一つ一つ聴くしかないため，時間を費やす必要があるからである．一方で，かわいさの評価の傾向は，人によって異なるとは言え，完全に異なるわけではなく，傾向に類似性が存在することも考えられる．

本章では，かわいさの評価の傾向には評価者間で類似性が存在すると仮定の下，協調フィルタリングを用いて，全音声に対して評価しなくてもかわいさの評価を推定し，そのユーザにとってかわいいと予想される話者を出力することを目指し，それがどの程度正しく推薦できるかを調査する．

4.2　協調フィルタリングによるかわいい話者の推薦

ユーザ（便宜上，新規評価者と呼ぶ）は予め M 人の話者集合の内 $M'(< M)$ 人の話者に対してかわいさを評価しているとする．このとき，M 個の評価値と新規

評価者以外のユーザ（参照評価者）の評価履歴を用いて，残り $M - M'$ 人の話者に対するかわいさの評価を予測する．

新規評価者 x による話者 i に対するかわいさの評価値を r_{xi}，参照評価者の集合を $\{\boldsymbol{U}\}$，評価者 $u \in \boldsymbol{U}$ による話者 i に対するかわいさの評価値を r_{ui} とする．このとき，協調フィルタリングは次の 3 段階の処理で実現される．

Phase I: 類似評価者の探索　各参照評価者 $u \in \boldsymbol{U}$ 毎に，新規評価者 x と参照評価者 u の類似度 W_{xu} を算出する．この類似度は両者の評価値対（r_{xi}, r_{ui}）を使用して計算する．また，類似度 W_{xu} は一般化 Jaccard 係数 [27]，相関係数，コサイン類似度などを用いる．各類似度の定義式はそれぞれ次の通りである．

$$J(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)} \tag{4.1}$$

$$R(\boldsymbol{x}, \boldsymbol{y}) = \frac{(\boldsymbol{x} - \overline{\boldsymbol{x}}) \cdot (\boldsymbol{y} - \overline{\boldsymbol{y}})}{\|\boldsymbol{x} - \overline{\boldsymbol{x}}\| \|\boldsymbol{y} - \overline{\boldsymbol{y}}\|} \tag{4.2}$$

$$C(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \tag{4.3}$$

このとき，ある θ_{sim} に対して，$W_{xu} > \theta_{\mathrm{sim}}$ なるような参照評価者 u を類似評価者と呼び，類似評価者集合を $\{\boldsymbol{U}_{\boldsymbol{x}}\}$ とする．

Phase II: かわいさ評価値の予測　新規評価者 x の未評価話者 i に対するかわいさの予測は次式で行う．

$$\hat{r}_{xi} = \begin{cases} \overline{\boldsymbol{r}_{\boldsymbol{x}}} + \dfrac{\sum_{u \in \boldsymbol{U}_{\boldsymbol{x}}} W_{xu}(r_{ui} - \overline{\boldsymbol{r}_u})}{\sum_{u \in \boldsymbol{U}_{\boldsymbol{x}}} |W_{xu}|} & (\boldsymbol{U}_{\boldsymbol{x}} \text{ is not empty}) \\ \overline{\boldsymbol{r}_{\boldsymbol{x}}} & (\boldsymbol{U}_{\boldsymbol{x}} \text{ is empty}) \end{cases} \tag{4.4}$$

ただし，$\overline{\boldsymbol{r}_{\boldsymbol{x}}}$ は r_{xi} の M' 話者分の平均値であり，$\overline{\boldsymbol{r}_u}$ は r_{ui} の M 話者分の平均値である．

Phase III: かわいい話者の推薦　$\hat{r}_{xi} \geq \theta_{\mathrm{rec}}$ を満たすとき，話者 i は新規評価者 x から θ_{rec} 以上の評価値を得られると見なし，話者 i を新規評価者 x に「かわいい話者」として推薦する．

4.3 実験条件

4.3.1 実験データ

音声データとして，ある声優事務所の公式ウェブページで公開されている 70 名分のサンプル音声の内，演技をせず声優自身の声色で自己紹介をしているものを使用した．各話者のサンプル音声から 5~7 秒の発話区間を 2 つずつ切り出して音声刺激として使用した．ただし，話者を一意に特定できる発話（例: 自身の名前を紹介している）や，笑い声などは除外した．

42 名（男性：31 名，女性：11 名，18~21 歳）に主観評価を行わせた．各評価者は，140 個の音声刺激を評価者毎に異なる順序で 1 つずつ聴取し（聴き直しは可），その都度「かわいいと思うか」を 7 段階（1：全くそう思わない ~ 7：とてもそう思う）で評価させた．

同一話者の 2 発話に対するそれぞれの評価値の平均値をその話者に対する評価とし，本実験では主観評価で得られた 42 名による 70 話者に対する評価値セットを実験データとして使用する．

4.3.2 分析条件

本実験は，leave-one-out アプローチによって，42 名の評価者から 1 名を選んで新規評価者 x とし，残り 41 名を参照評価者 $u \in \{\boldsymbol{U}_{\boldsymbol{x}}\}$ とする．新規評価者 x の評価値の内，ランダムに選んだ 60 話者分を未評価と見なし，10 話者分の評価値と，参照評価者の 70 話者分の評価値を用いて推定する．また，推定値に基づいてかわいい話者を新規評価者 x に推薦する．予測精度として，Mean Absolute Eorrors (MAEs), 推薦精度として，F 値，適合率および再現率を算出する．また，予測結果は予測対象とする話者 60 名の組み合わせに依存するため，各新規評価者 x につき 1000 回行い，予測精度及び推薦制度はその 1000 回分の精度の平均とする．θ_{sim} は 0.25, 0.50 および 0.75 を用い，θ_{rec} は 5.0 とする．

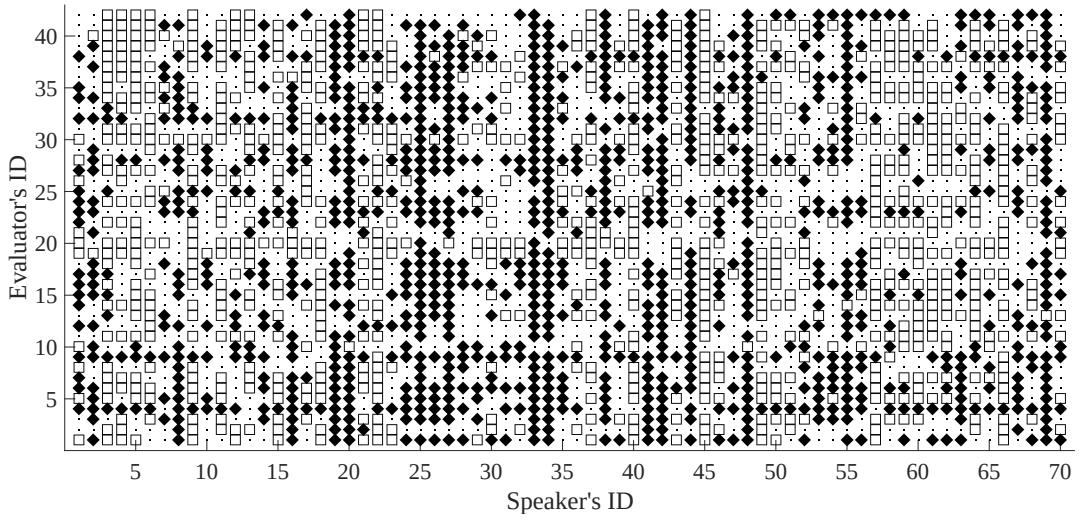

図 4.1: かわいさの主観評価結果 (◆:1 ≤ value ≤ 3　·:3 < value < 5　□:5 ≤ value ≤ 7)

表 4.1: MAEs の平均値 (標準偏差)

	.25	.50	.75
Jaccard index	.865 (.211)	.865 (.210)	.854 (.231)
Correlation	.855 (.206)	.856 (.206)	.888 (.212)
Cosine	.867 (.211)	.867 (.211)	.868 (.211)

4.4　実験結果・考察

各被験者による全話者に対する評価結果を図 4.1 に，予測結果として MAEs を表 4.1 に示す． MAEs はどれも 0.9 未満であった．また，推薦精度を表 4.2 に示す．ただし，4 番目の新規評価者については，どの話者にも 5 以上の評価を与えなかったため，表 4.2 の算出からは除外した．F 値の平均値及び中央値は 0.4 ∼ 0.6 だったが，適合率は高かった（平均が 0.7 前後，中央値が 0.8 前後）．かわいい話者を推薦するという目的においては，適合率は特に重要な指標なため，かわいい話者の推薦に対して本手法の有効性が見られる．次に，F 値が上位 25% 及び下位 25% だった新規評価者における精度を表 4.3 に示す．F 値が上位 25% の場合は（ID:20

表 4.2: 推薦制度の平均値（標準偏差）と中央値（四分位範囲）

(a) F-score

	Mean			Median		
	.25	.50	.75	.25	.50	.75
Jaccard	.460 (.188)	.461 (.187)	.504 (.186)	.444 (.363)	.446 (.356)	.477 (.357)
Correlation	.495 (.183)	.507 (.182)	.523 (.184)	.496 (.322)	.498 (.324)	.543 (.355)
Cosine	.459 (.185)	.457 (.187)	.458 (.188)	.445 (.338)	.440 (.337)	.445 (.349)

(b) Recall rate

	Mean			Median		
	.25	.50	.75	.25	.50	.75
Jaccard	.376 (.219)	.375 (.220)	.409 (.225)	.412 (.382)	.411 (.387)	.398 (.424)
Correlation	.416 (.218)	.425 (.217)	.420 (.225)	.470 (.370)	.448 (.379)	.453 (.400)
Cosine	.372 (.217)	.372 (.218)	.371 (.220)	.404 (.368)	.401 (.370)	.405 (.378)

(c) Precision rate

	Mean			Median		
	.25	.50	.75	.25	.50	.75
Jaccard	.736 (.251)	.732 (.246)	.693 (.245)	.844 (.202)	.848 (.202)	.791 (.320)
Correlation	.720 (.245)	.710 (.241)	.671 (.241)	.845 (.234)	.833 (.267)	.791 (.387)
Cosine	.732 (.248)	.735 (.249)	.734 (.248)	.842 (.204)	.844 (.209)	.845 (.210)

の新規評価者を除いて）類似評価者数が 10 人を超えており，一方で下位 25% の場合は（ID:18, 29 の新規評価者を除いて）10 人未満であった．また，ID:10, 25 の新規評価者については類似評価者数の平均が 1 未満であり，F 値が特に低かった（$F < 0.2$）．一般的に，協調フィルタリングによる推薦精度は類似評価者の人数が影響すると言われており，本実験の結果と一致する．

表 4.3 中の新規評価者に関して，図 4.1 から，ID:20, 30 の新規評価者は 60% 以上の話者に対してかわいさを知覚しており（score ≤ 5），ID:19, 24, 37 では，その割合が 50 $\sim$ 60% であった．一方で，ID:10, 12, 21, 25 は 10 $\sim$ 20% の話者にしかかわいさを知覚していなかった．また，5 以上の評価を付けた話者数と類似評価者数との相関係数は 0.649 であった．従って，話者に対してかわいさを知覚しにくい場合，その新規評価者は類似評価者が少ない可能性が高く，協調フィルタリン

表 4.3: F 値が上位 25%, 下位 25% のユーザに対する推薦結果
(a) Top 25%

User's ID	11	14	19	20	22	24	30	31	37	40
Mean of #similar	16.2	13.6	10.3	9.1	15.6	17.2	11.5	15.3	16.7	14.8
MAE	.629	.708	.812	.704	.690	.766	.766	.725	1.092	.882
F-value	.745	.751	.748	.800	.736	.740	.773	.713	.755	.720
Precision	.862	.893	.826	.846	.862	.964	.889	.877	.909	.831
Recall	.702	.637	.690	.760	.684	.643	.698	.662	.698	.680

(b) Bottom 25%

User's ID	10	12	18	21	25	29	32	33	36	38
Mean of #similar	.2	5.2	14.0	2.2	.7	10.0	3.3	4.6	5.8	1.5
MAE	.961	.914	.680	.818	1.015	.878	1.309	1.023	.914	1.860
F-value	.170	.341	.338	.324	.190	.345	.307	.284	.318	.245
Precision	.182	.373	.605	.265	.134	.572	.567	.317	.435	.354
Recall	.015	.267	.260	.220	.046	.243	.127	.195	.225	.062

グによる評価値の予測・話者の推薦が難しいと考えられる.

男女の差について検討する．男性新規評価者に対して，類似評価者として選ばれた評価者の内，男性と女性の比は 2.304 : 1.000 であり，女性新規評価者に対しては 2.807 : 1.000 であった．この 2 つの比に大きな差がなく，本実験では男女間で評価の傾向に大きな差は見られなかった．文献 [6] でも，女性声における魅力の知覚に男女間で正の相関があると述べており，本結果と一致する.

4.5　本章のまとめ

本章では，ユーザ間でかわいさの評価の傾向に類似性があると仮定し，協調フィルタリングによる評価値の予測及びかわいい話者の推薦を行った.

(1) 評価値の予測精度である MAEs は 0.9 程度で，話者推薦における適合率の中央値は 0.8 であった.
(2) 高評価を与えた話者の数と類似評価者数に正の相関があった（相関係数:0.649）.

(3) かわいさの評価の傾向において，男女間で大きな差は見られなかった．

その結果から，ユーザ間でかわいさの評価の傾向に類似性が存在するという仮定の下，かわいさを感じる話者が少ないユーザを除き，未知評価話者に対する評価値の予測及びかわいい話者の推薦における有効性が示された．

第5章　話者適応型Restricted Boltzmann Machineに基づくかわいさの制御

3章では，大局的特徴を操作せずとも，局所的特徴だけでもかわいさを制御できるという結果が得られ，局所的特徴（スペクトル）の変換による声質変換でかわいさの制御を行えるという可能性を示した．4章では，ユーザの少数の話者に対するかわいさの評価結果から，そのユーザの多数の話者に対する評価値を推定できることが示唆された．

本章では，ユーザの多数の話者に対するかわいさの評価結果を下に，スペクトル変換による声質変換を用いたかわいさの制御方法を提案する．

5.1　はじめに

従来の声質変換によって，ユーザがかわいいと思う話者を目標話者に設定することでかわいい音声を得ることができる．しかし，出力されるスペクトルの特徴は同一人物のものに限定されてしまい，ある一人の人物のものとして出力してしまうのは応用の幅を限定してしまう．入力音声をかわいく加工した音声を出力したい場合，変換対象であるスペクトルの内，かわいさに起因する情報とそれ以外の情報に分離可能なモデルを考える必要がある．それによって，かわいさに起因する情報のみをユーザの思うかわいい話者のものに変えることで，かわいく加工された音声が生成できると期待できる．

本章では，話者に起因する情報とそれ以外の情報に分離可能なモデルである話者適応型 Restricted Boltzmann Machine[21] を拡張し，かわいさに起因する情報

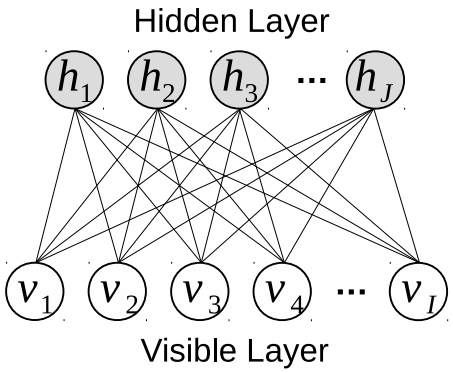

図 5.1: RBM のモデル図

も分離可能なモデルを提案する．

5.2 Gaussian-Bernoulli 型 Restricted Boltzmann Machine

Restricted Boltzmann Machine(RBM)[28] は可視層と隠れ層の 2 つの層からなる無向グラフィカルモデルである（図 5.1）．RBM の 2 つの層に対応する特徴量はどちらもベルヌーイ分布に従うが，Gaussian-Bernoulli 型の RBM（GBRBM）は可視層が正規分布に従う．可視層と隠れ層の特徴量それぞれが $\boldsymbol{v} = [v_1, \cdots, v_i, \cdots, v_I]^\top$, $v_i \in \mathbb{R}$, $\boldsymbol{h} = [h_1, \cdots, h_j, \cdots, h_J]^\top$, $h_j \in \{0, 1\}$ で，可視層の分散が $\boldsymbol{\sigma}^2 = [\sigma_1^2, \cdots, \sigma_i^2, \cdots, \sigma_I^2]^\top$, $\sigma_i^2 \in \mathbb{R}$ とするとき，GBRBM のエネルギー関数，分配関数及び $\boldsymbol{v}$ と $\boldsymbol{h}$ の同時確率 $p(\boldsymbol{v}, \boldsymbol{h})$ はそれぞれ次のように定義される．

$$E(\boldsymbol{v}, \boldsymbol{h}) = \left|\left|\frac{\boldsymbol{v} - \boldsymbol{b}}{2\boldsymbol{\sigma}}\right|\right|^2 - \boldsymbol{c}^\top \boldsymbol{h} - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right)^\top \boldsymbol{W}\boldsymbol{h} \tag{5.1}$$

$$Z = \sum_{\boldsymbol{v}, \boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \tag{5.2}$$

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \tag{5.3}$$

ただし，$\boldsymbol{W} \in \mathbb{R}^{I \times J}$, $\boldsymbol{b} \in \mathbb{R}^{I \times 1}$, 及び $\boldsymbol{c} \in \mathbb{R}^{J \times 1}$ はそれぞれ可視層と隠れ層間の接続重み行列，可視層のバイアス及び隠れ層のバイアスである．また，$||\cdot||^2$ は L2 ノル

ム，$\frac{\boldsymbol{x}}{\boldsymbol{y}}$ はベクトル $\boldsymbol{x}, \boldsymbol{y}$ の要素除算を表す．GBRBM では（RBM も同様）可視層のユニット間及び隠れ層のユニット間には接続がない（つまり条件付き独立である）ため，条件付き確率 $p(\boldsymbol{h}|\boldsymbol{v}), p(\boldsymbol{v}|\boldsymbol{h})$ はそれぞれ次のようになる．

$$p(h_j = 1|\boldsymbol{v}) = \text{sigmoid}\left(c_j + \sum_i W_{ij}\frac{v_i}{\sigma_i^2}\right) \tag{5.4}$$

$$p(v_i = v|\boldsymbol{h}) = \mathcal{N}\left(v; b_i + \sum_j W_{ij}h_j, \sigma_i^2\right) \tag{5.5}$$

sigmoid$(\cdot)$ は要素毎のシグモイド関数，$\mathcal{N}(\cdot;\ \mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布である．

GBRBM のパラメータ $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\sigma}^2\}$ は対数尤度 $\mathcal{L} = \log\prod_n p(\boldsymbol{v}_n)$ が最大となるように最急降下法で更新される．また，GBRBM では分散 $\boldsymbol{\sigma}^2$ を非負値に制約する必要があり，その学習を安定させるために $\boldsymbol{z} = [z_1, \cdots, z_i, \cdots, z_I]^\top$, $z_i = \log\sigma_i^2$ を $\boldsymbol{\sigma}^2$ の代わりに学習する．

5.3 話者適応型 RBM

5.3.1 概要

話者適応型 RBM[21] (ARBM) は GBRBM を拡張した, Fig. 5.2 のような無向グラフィカルモデルである．ARBM では，話者に依存する情報と話者に依存しない情報を分離しながら，潜在的な特徴を抽出するような確率モデルである．

可視層の音響特徴量 $\boldsymbol{v}$ と隠れ層の潜在特徴量 $\boldsymbol{h}$(one-hot vector) 間の対応を話者識別ベクトル $\boldsymbol{s} = [s_1, \cdots, s_r, \cdots, s_R]^\top]$ （話者 r に対して $s_r = 1$ となる one-hot vector．R は参照話者数）で制御される話者依存パラメータと話者不定パラメータで表現できると仮定し，以下の式で確率密度関数を定義する．

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{Z}e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{s})} \tag{5.6}$$

$$E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{2}\left\|\frac{\boldsymbol{v} - \boldsymbol{b}(\boldsymbol{s})}{\boldsymbol{\sigma}}\right\|^2 - \boldsymbol{c}(\boldsymbol{s})^\top\boldsymbol{h} - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}\right)^\top\boldsymbol{W}(\boldsymbol{s})\boldsymbol{h} \tag{5.7}$$

$$Z = \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{s})} \tag{5.8}$$

ただし

$$\boldsymbol{W}(\boldsymbol{s}) = \sum_r \boldsymbol{A}_r s_r \overline{\boldsymbol{W}} \tag{5.9}$$

$$\boldsymbol{b}(\boldsymbol{s}) = \overline{\boldsymbol{b}} + \sum_r \boldsymbol{b}_r s_r = \overline{\boldsymbol{b}} + \boldsymbol{B}\boldsymbol{s} \tag{5.10}$$

$$\boldsymbol{c}(\boldsymbol{s}) = \overline{\boldsymbol{c}} + \sum_r \boldsymbol{c}_r s_r = \overline{\boldsymbol{c}} + \boldsymbol{C}\boldsymbol{s} \tag{5.11}$$

$\overline{\boldsymbol{W}} \in \mathbb{R}^{I\times J}$, $\overline{\boldsymbol{b}} \in \mathbb{R}^{I\times 1}$, $\overline{\boldsymbol{c}} \in \mathbb{R}^{J\times 1}$ は話者不定パラメータであり，$\boldsymbol{A}_r \in \mathbb{R}^{I\times I}$ と $\boldsymbol{b}_r \in \mathbb{R}^{I\times 1}$ $\left(\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_I] \in \mathbb{R}^{I\times R}\right)$ 及び $\boldsymbol{c}_r \in \mathbb{R}^{J\times 1}$ $\left(\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_J] \in \mathbb{R}^{J\times R}\right)$ は話者依存パラメータである．また，可視層の分散 $\boldsymbol{\sigma}^2 \in \mathbb{R}^{I\times 1}$ は GBRBM と同様であり，ここでは話者不定パラメータとして扱う．また，両層の条件付き確率は次の式になる．

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{s}) = \mathcal{S}\left(c_j(\boldsymbol{s}) + \sum_i W_{ij}(\boldsymbol{s})\frac{v_i}{\sigma_i^2}\right) \tag{5.12}$$

$$p(v_i = v|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}\left(v; b_i(\boldsymbol{s}) + \sum_j W_{ij}(\boldsymbol{s})h_j, \sigma_i^2\right) \tag{5.13}$$

$\mathcal{S}(\cdot)$ は要素ごとのソフトマックス関数（式 5.14）である．

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{s}) = \frac{e^{c_j(\boldsymbol{s}) + \sum_i W_{ij}(\boldsymbol{s})v_i'}}{\sum_k e^{c_k(\boldsymbol{s}) + \sum_i W_{ik}(\boldsymbol{s})v_i'}} \quad (v_i' = \frac{v_i}{\sigma_i^2}) \tag{5.14}$$

5.3.2　学習・適応プロセス

ARBM のパラメータ $\boldsymbol{\theta} = \{\overline{\boldsymbol{W}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, \boldsymbol{\sigma}^2, \{\boldsymbol{A}_r\}_{r=1}^R, \boldsymbol{B}, \boldsymbol{C}\}$ は対数尤度 $\mathcal{L} = \log\prod_n p(\boldsymbol{v}_n|\boldsymbol{s})$ が最大となるように最急降下法によって推定する．

新規の話者をモデルに適応したい場合，既に最適化し終わった話者不定パラメータを初期値として固定したまま，その新規話者の少量のデータを用いて当該話者の依存パラメータのみを学習時同様に推定する．このとき，参照話者数は R から $R+1$ に増える．

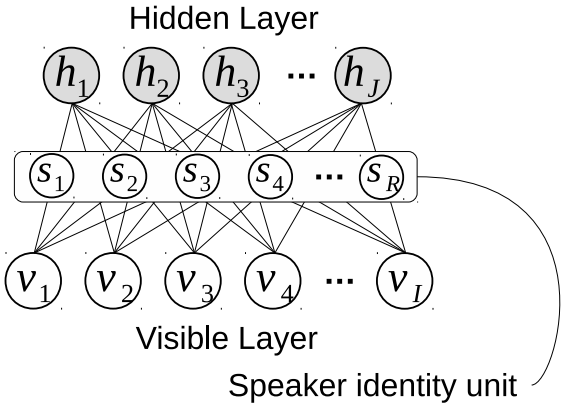

図 5.2: ARBM のモデル図

5.3.3 変換プロセス

元話者の音声特徴量 $\boldsymbol{v}_x^{(t)}$ から目標話者の音声特徴量 $\hat{\boldsymbol{v}}_y^{(t)}$ を，$p(\boldsymbol{v}_y^{(t)}|\boldsymbol{v}_x^{(t)})$ が最大となるように推定する．つまり，

$$
\begin{aligned}
\hat{\boldsymbol{v}}_y^{(t)} &\triangleq \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{v}_y^{(t)}|\boldsymbol{v}_x^{(t)}) \\
&= \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}} \sum_{\boldsymbol{h}^{(t)}} p(\boldsymbol{h}^{(t)}|\boldsymbol{v}_x^{(t)}) p(\boldsymbol{v}_y^{(t)}|\boldsymbol{h}^{(t)}) \\
&\simeq \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{h}}^{(t)}|\boldsymbol{v}_x^{(t)}) p(\boldsymbol{v}_y^{(t)}|\hat{\boldsymbol{h}}^{(t)}) \\
&= \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{v}_y^{(t)}|\hat{\boldsymbol{h}}^{(t)}) \\
&= \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, \mathcal{N}(\boldsymbol{v}_y^{(t)};\ \overline{\boldsymbol{b}} + \boldsymbol{b}_y + \boldsymbol{A}_y \overline{\boldsymbol{W}} \hat{\boldsymbol{h}}^{(t)}, \boldsymbol{\sigma}^2) \\
&= \overline{\boldsymbol{b}} + \boldsymbol{b}_y + \boldsymbol{A}_y \overline{\boldsymbol{W}} \hat{\boldsymbol{h}}^{(t)} \qquad (5.15) \\
\hat{\boldsymbol{h}}^{(t)} &\triangleq \underset{\boldsymbol{h}^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{h}^{(t)}|\boldsymbol{v}_x^{(t)}) \\
&\simeq \mathbb{E}[\boldsymbol{h}^{(t)}|\boldsymbol{v}_x^{(t)}] \\
&= \mathcal{S}\left(\overline{\boldsymbol{c}} + \boldsymbol{c}_x + \overline{\boldsymbol{W}}^{\top} \boldsymbol{A}_x^{\top} \left(\frac{\boldsymbol{v}_x^{(t)}}{\boldsymbol{\sigma}^2}\right)\right) \qquad (5.16)
\end{aligned}
$$

5.4　ARBM に基づくかわいさの制御法

5.4.1　概要

話者の個人性にはかわいい (あるいはかわいくない) 話者間で共通な項が含まれていると仮定し，話者依存パラメータとその共通項を分けて扱う. 今回は予めユーザがかわいいと判断した話者 (クラス K) とかわいくないと判断した話者 (クラス $\overline{\mathrm{K}}$) の 2 つのクラスを考え，話者識別ベクトル $\boldsymbol{s}$ とかわいさクラスの識別ベクトル $\boldsymbol{d} = [d_{\mathrm{K}}, d_{\overline{\mathrm{K}}}]^{\top}$ （話者識別ベクトル $\boldsymbol{s}$ の話者の属するクラスにより定まる one-hot vector）の 2 つで射影を制御する．これにより，潜在特徴量から音響特徴量を復元する際に，クラス依存パラメータを適切に切り替えることで所望するクラスの要素を持った音声が生成できることが期待できる．このとき，両層間の制約（式 5.9, 5.10, 5.11）は次式に変わる.

$$\boldsymbol{W}(\boldsymbol{s}) = \sum_{k}\sum_{r} \boldsymbol{H}_k d_k \boldsymbol{A}_r s_r \overline{\boldsymbol{W}} \tag{5.17}$$

$$\boldsymbol{b}(\boldsymbol{s}) = \overline{\boldsymbol{b}} + \sum_{k} \boldsymbol{b}'_k d_k + \sum_{r} \boldsymbol{b}_r s_r = \overline{\boldsymbol{b}} + \boldsymbol{B}'\boldsymbol{d} + \boldsymbol{B}\boldsymbol{s} \tag{5.18}$$

$$\boldsymbol{c}(\boldsymbol{s}) = \overline{\boldsymbol{c}} + \sum_{k} \boldsymbol{c}'_k d_k + \sum_{r} \boldsymbol{c}_r s_r = \overline{\boldsymbol{c}} + \boldsymbol{C}'\boldsymbol{d} + \boldsymbol{C}\boldsymbol{s} \tag{5.19}$$

$\boldsymbol{H}_k \in \mathbb{R}^{I\times I}$，$\boldsymbol{b}'_k \in \mathbb{R}^{I\times 1}$ $\left(\boldsymbol{B}' = [\boldsymbol{b}'_{\mathrm{K}}, \boldsymbol{b}'_{\overline{\mathrm{K}}'}]\right)$ 及び $\boldsymbol{c}'_k \in \mathbb{R}^{J\times 1}$ $\left(\boldsymbol{C}' = [\boldsymbol{c}'_{\mathrm{K}}, \boldsymbol{c}'_{\overline{\mathrm{K}}'}]\right)$ はかわいさクラス依存パラメータである.

5.4.2　学習・適応プロセス

本モデルで推定すべきパラメータは $\boldsymbol{\theta} = \{\overline{\boldsymbol{W}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, \boldsymbol{\sigma}^2, \{\boldsymbol{A}_r\}_{r=1}^{R}, \boldsymbol{B}, \boldsymbol{C}, \{\boldsymbol{H}_k\}_k, \boldsymbol{B}', \boldsymbol{C}'\}$ である．ARBM は学習に用いる話者数が多くなると話者不定パラメータの推定が不安定になる可能性がある [21]．ユーザによってクラス K 及び $\overline{\mathrm{K}}$ の話者数は多くなる場合があるため，話者不定パラメータの推定が不安定となる可能性がある．従って，学習・適応を 3 段階に分けて行う.

(1) まず，少数の話者の十分な量の学習データで従来の ARBM を学習する．これにより，話者不定パラメータの推定が終了したこととなる.

(2) 次に，2 種のクラスの話者のデータを用いてクラス依存パラメータのみを最適化する．ただしこのとき，話者不定パラメータは (1) の最終状態のもので固定しておく．

(3) 最後に，2 種のクラスの話者のデータを用いて，クラス依存パラメータと共に各話者の話者依存パラメータを最適化する．

(2) でクラス依存パラメータのみを最適化する理由は，話者依存パラメータと同時に最適化を行ったときにクラス依存パラメータに吸収されるべき成分が話者依存パラメータに吸収されることを防ぐためである．先にクラス依存パラメータのみで最適化しておき，そのクラス依存パラメータだけでは表現できない項は話者依存項であると見なし，話者依存パラメータに吸収させる．

5.4.3 変換プロセス

元話者の音声特徴量 $\boldsymbol{v}_x^{(t)}$ から目標話者の音声特徴量 $\hat{\boldsymbol{v}}_y^{(t)}$ は次式で推定する．

$$
\begin{aligned}
\hat{\boldsymbol{v}}_y^{(t)} &\triangleq \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{v}_y^{(t)}|\boldsymbol{v}_x^{(t)}) \\
&= \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}} \sum_{\boldsymbol{h}^{(t)}} p(\boldsymbol{h}^{(t)}|\boldsymbol{v}_x^{(t)}) p(\boldsymbol{v}_y^{(t)}|\boldsymbol{h}^{(t)}) \\
&\simeq \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{h}}^{(t)}|\boldsymbol{v}_x^{(t)}) p(\boldsymbol{v}_y^{(t)}|\hat{\boldsymbol{h}}^{(t)}) \\
&= \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{v}_y^{(t)}|\hat{\boldsymbol{h}}^{(t)}) \\
&= \underset{\boldsymbol{v}_y^{(t)}}{\operatorname{argmax}}\, \mathcal{N}(\boldsymbol{v}_y^{(t)};\ \overline{\boldsymbol{b}} + \boldsymbol{b}'_y + \boldsymbol{b}_y + \boldsymbol{H}_y \boldsymbol{A}_y \overline{\boldsymbol{W}} \hat{\boldsymbol{h}}^{(t)}, \boldsymbol{\sigma}^2) \\
&= \overline{\boldsymbol{b}} + \boldsymbol{b}' + \boldsymbol{b}_y + \boldsymbol{H}_y \boldsymbol{A}_y \overline{\boldsymbol{W}} \hat{\boldsymbol{h}}^{(t)} \qquad (5.20) \\
\hat{\boldsymbol{h}}^{(t)} &\triangleq \underset{\boldsymbol{h}^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{h}^{(t)}|\boldsymbol{v}_x^{(t)}) \\
&\simeq \mathbb{E}[\boldsymbol{h}^{(t)}|\boldsymbol{v}_x^{(t)}] \\
&= \mathcal{S}\left(\overline{\boldsymbol{c}} + \boldsymbol{c}'_x + \boldsymbol{c}_x + \overline{\boldsymbol{W}}^\top \boldsymbol{A}_x^\top \boldsymbol{H}_x^\top \left(\frac{\boldsymbol{v}_x^{(t)}}{\boldsymbol{\sigma}^2} \right) \right) \qquad (5.21)
\end{aligned}
$$

また，かわいさを個別に制御する場合，元話者がクラス $\overline{\mathrm{K}}$（かわいくない話者）に属するとすると

$$\hat{\boldsymbol{v}}_{x'}^{(t)} \triangleq \overline{\boldsymbol{b}} + \boldsymbol{b}'_{\mathrm{K}} + \boldsymbol{b}_x + \boldsymbol{H}_{\mathrm{K}} \boldsymbol{A}_x \overline{\boldsymbol{W}} \hat{\boldsymbol{h}}^{(t)} \tag{5.22}$$

$$\hat{\boldsymbol{h}}^{(t)} \triangleq \mathcal{S}\left(\overline{\boldsymbol{c}} + \boldsymbol{c}'_{\overline{\mathrm{K}}} + \boldsymbol{c}_x + \overline{\boldsymbol{W}}^{\top} \boldsymbol{A}_x^{\top} \boldsymbol{H}_{\overline{\mathrm{K}}}^{\top} \left(\frac{\boldsymbol{v}_x^{(t)}}{\boldsymbol{\sigma}^2}\right)\right) \tag{5.23}$$

このように，話者依存パラメータは元話者のものを使用し，クラス依存パラメータのみクラス K のものに差し替えることで実現される．

5.5　実験的評価

5.5.1　実験条件

提案法を用いたかわいさの制御によって音声のかわいさが高くなる事を示すため，日本人男性 12 名（平均：21.8 歳，標準偏差：1.1 歳）に対して主観評価実験を行った．各被験者は事前に自らがかわいくないと評価したクラス $\overline{\mathrm{K}}$ の話者の音声を次の 3 手法で変換した音声を評価する．

X2Y: クラス K の話者へ変換する（式 5.20, 5.21 に従う．ARBM-VC と等価）
X2X': 入力話者とは異なるクラス $\overline{\mathrm{K}}$ の話者へ変換する（式 5.20, 5.21 に従う．ARBM-VC と等価）
PRO: 入力話者とは異なるクラス $\overline{\mathrm{K}}$ の話者へ変換する．ただし，クラス依存パラメータを K のものに差し替える（式 5.22, 5.23 に従う．提案法）

ただし，**X2X'** と **PRO** の変換目標の話者は同一とし，元話者と目標話者の組み合わせはランダムに 10 組選んだ．評価項目は

かわいさ： かわいさについて 3 種の音声の順位付け
かわいさ： 提案法 **PRO** の生成音声のかわいさ (1:かわいくない，7:かわいい)
話者性: **PRO** の音声が **X2Y** と **X2X'** の音声のどちらに近いか（1:**X2Y**，7:**X2X'**）

とした．つまり，本実験では **PRO** の音声が順位付けで上位に入り，そのかわいさが変換前の評価よりも高くなることを示す．また，提案法のかわいさの制御による話者性への影響も確認する．

事前評価として日本人女性声優 50 名分のボイスサンプルを聴かせ，各話者のかわいさを 7 段階で評価させた（1:かわいくない，7:かわいい）．ユーザ毎に，6 以上の評価を得た話者をクラス K に，2 以下をクラス $\overline{\mathrm{K}}$ とした．ARBM の学習用音声データには独自に収録した日本人女性話者 8 名による音素バランス文読み上げ音声（各話者 46 ∼ 50 発話）を用いた．クラス依存・話者依存パラメータの学習，及び声質変換には 2 クラス分の話者のボイスサンプルを使用した（その他の評価の話者は使用しない）．

分析合成には WORLD (D4C edition) [24, 25] を用い，音響特徴量には 24 次のメルケプストラムを使用した．モデルの可視層は 24 次元，隠れ層は 8 次元とし，モデルの学習は学習率 0.01，モーメンタム 0.9，バッチサイズ $100 \times R$ とする．ただし，ARBM の学習（5.4.2 節 (1)）では $R = 8$，クラス依存・話者依存パラメータの学習（5.4.2 節 (2)(3)）では R は被験者毎に異なる．また，学習のエポックは ARBM の学習（5.4.2 節 (1)）は 100，クラス依存パラメータの学習（5.4.2 節 (2)）は 100，クラス依存パラメータと話者依存パラメータの同時学習 (5.4.2 節 (3)) は 50 回とし，確率的勾配法で学習した．また，生成音声の変調スペクトルをポストフィルタで補償した [30].

5.5.2 実験結果

図 5.3 に順位付けの結果を示す．(a) 及び (b) から，**X2Y** よりも **X2X'** による生成音声の方が若干かわいさが高かった．これは，**X2Y** は **X2X'** とは違い，異なるクラスの話者間での変換だったため話者間の類似性が低く，不自然な音声を生成してしまったという場合が考えられる．一方で，被験者が事前にかわいくないと評価した話者へ変換しているにも関わらず，**PRO** の生成音声はその 80% 程度が 2 以内に順位付けられており，また，かわいさの評価の半数以上が 3 を超えていた（平均:3.1, 標準偏差:1.5, 図 5.4）．これは，目標話者もクラス $\overline{\mathrm{K}}$ から選んでいるため，**X2Y** よりも類似性の近い話者同士の変換であったために不自然さを抑えられたことと，かわいさ依存の成分を制御したことが要因と考えられる．

従って，クラス依存パラメータではかわいさに影響する成分を表現できていると言える．

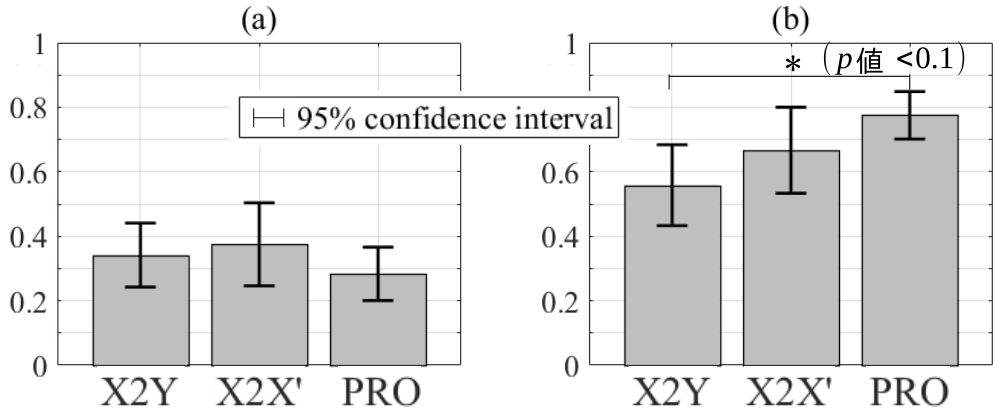

図 5.3: 順位付けにおいて，(a)1 位として選択された割合 (b)1 位あるいは 2 位として選択された割合

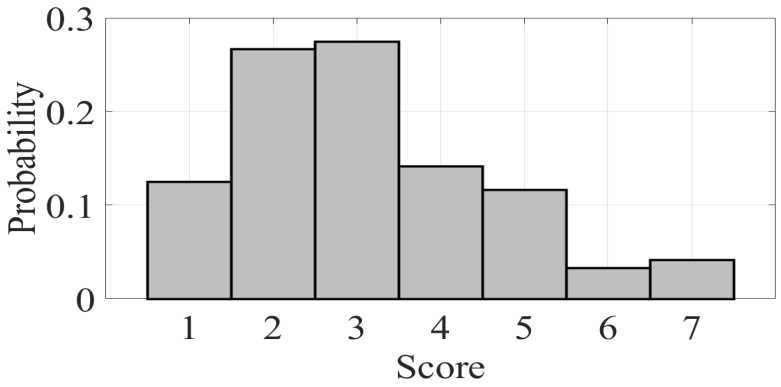

図 5.4: かわいさのスコアの分布

また，話者性に関するテストの評価値分布については図 5.5 から，**X2X'** と同じ目標話者に変換しているにも関わらず，**PRO** による生成音声の話者性はその 44.2% 程が **X2Y** のものに近かった．同じ話者を目標話者としていても，クラス依存パラメータを制御したことで出力話者の話者性が大きく変化する場合があるということを示している．

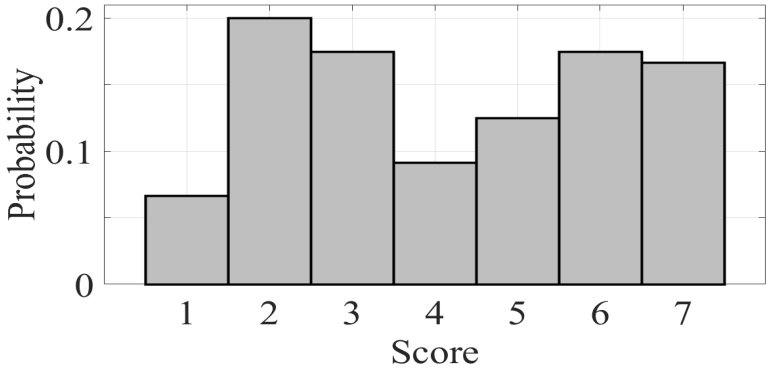

図 5.5: 話者性に関するスコアの分布

5.6 本章のまとめ

本章では，音声における「かわいさ」の制御を目指し，話者適応型 RBM を拡張し，その学習法を提案した．主観評価の結果，提案法は，かわいくない話者（かわいさの主観評価値が 2 以下）へ変換しているにも関わらず，

(1) 従来の声質変換手法と比較して，80% 近くの回答でかわいさの順位が上位だった．
(2) 主観評価値が平均で 3 以上で，60% 以上が 3 以上の評価だった．
(3) 話者性がかわいいクラスの話者に近いという回答がおよそ 44.2% だった．

また，話者性の制御とは別にかわいさの制御を可能にすることで，元話者に近い話者を目標話者に設定することができるようになった．これにより，元話者と目標話者の話者空間上での距離が遠い場合に起こる音質劣化の影響を低減できると期待できる．

第6章　結論

6.1　おわりに

本論文ではユーザの思うかわいい声の生成を実現するため，

(1) かわいさの評価と聴取時間の関係性
(2) ユーザが思うかわいい話者の推薦
(3) かわいさのみの制御

この3つの課題に着目した．

3章では，音声におけるかわいさの評価と，聴取時間の長さとの関係性を調査した．その結果，聴取時間が短い場合，音声のかわいさに対する評価の安定性は低くなるものの，75 msの音声に対してもかわいさの評価を行うことは可能であり，スペクトルの変換による声質変換によってかわいさの制御ができることが示唆された．

4章では，ユーザ間でかわいさの評価の傾向に類似性があるという仮説の元，協調フィルタリングによる評価値の予測及びかわいい話者の推薦を行い，それがどの程度正しく推薦できるかのかを調査した．その結果から，協調フィルタリングは，かわいさを感じる話者が少ないユーザを除き，未知評価話者に対する評価値の予測及びかわいい話者の推薦に有効であることが示唆された．

5章では，音声における「かわいさ」の制御を目指し，話者適応型RBMを拡張したモデルとその学習法を提案した．提案法では，かわいくない話者へ変換しているにも関わらず，かわいさに影響する特徴のみを制御したことで，変換前よりもかわいさが向上し，さらに，従来の声質変換手法よりもかわいさが優位であった．

6.2　今後の課題

今後の課題について述べる．

まず，4 章における，かわいい話者の推薦について，適合率が 7 割程度であった．本研究で目指している声質変換システムでは，ユーザの思うかわいい話者とそうでない話者を適切に分類することが必要とされているため，7 割という推薦精度は不十分である．今回の 4 章の実験では，評価者数を増やすことと，AutoEncoder や RBM などのディープラーニングを用いた推薦アルゴリズム [31, 32] を適用することで，精度の向上が期待される．

5 章における声質変換では，生成音声に不自然な音声が含まれており，その不自然さがかわいさを不当に低減させている要因となっていた．本手法（及び ARBM）では，潜在特徴に話者に依存しない特徴として，音韻特徴が抽出されることを期待している．しかし，パラメータの学習は,（入力音声から得られた）潜在特徴から入力音声の音響特徴量の分布を推定するような基準になっているため，潜在特徴には話者に依存する特徴が多少含まれている可能性がある．従って，学習において，話者の情報を変換するという明示的な条件を設ける必要がある．

また，本研究では，ユーザの思うかわいい話者とそうでない話者の情報に基づいた声質変換システムの提案と，その情報を得るための話者に対するかわいさの評価値の推定を行った．しかし，推定の誤りが変換音声のかわいさにどれだけの影響を与えるかは調査できておらず，今後の課題である．

参考文献

[1] Hill, A. K., and Puts, D. A.: "Vocal Attractiveness," *Encyclopedia of Evolutionary Psychological Science*, Springer International Publishing, 2016.

[2] Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., and Vukovic, J.: "Integrating cues of social interest and voice pitch in men's preferences for women's voices," *Biology Letters*, vol. 4, no. 2, pp. 192–194, 2008.

[3] Liu, X., and Xu, Y.: "What makes a female voice attractive," *Proceedings of XVIIth ICPHS*, Hong Kong, pp. 1274–1277, 2011.

[4] Xu, Y., Lee, A., Wu, W. L., Liu, X., and Birkholz, P.: "Human vocal attractiveness as signaled by body size projection," *PloS one*, vol. 8, no. 4, e62397, 2013.

[5] Ferdenzi, C., Patel, S., Mehu-Blantar, I., Khidasheli, M., Sander, D., and Delplanque, S.: "Voice attractiveness: Influence of stimulus duration and type," *Behavior research methods*, vol. 45, no. 2, pp. 405–413, 2013.

[6] Babel, M., McGuire, G., and King, J.: "Towards a more nuanced view of vocal attractiveness," *PloS one*, vol. 9, no. 2, e88616, 2014.

[7] Puts, D. A., Barndt, J. L., Welling, L. L., Dawood, K., and Burriss, R. P.: "Intrasexual competition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness," *Personality and Individual Differences*, vol. 50, no. 1, pp. 111–115, 2011.

[8] 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、STRAIGHT を用いた合成音声評価,” 日本音響学会春季講演論文集, pp. 503–506, 2014.

[9] Takano, S., and Yamada, M.: “Perception and analysis of “Moe” voice,” *The Journal of the Acoustical Society of America*, vol. 140, no. 4, p. 3399–3399, 2016.

[10] Kawahara, S.: “The phonetics of Japanese maid voice I: Apreliminary study,” *Phonological Studies*, pp. 19–28, 2013.

[11] Kawahara, S.: “The Prosodic Features of the “moe” and “tsun” Voices,” *Journal of the Phonetic Society of Japan*, vol. 20, no. 2, pp. 102–110, 2016.

[12] Starr, R. L.: “Sweet voice: The role of voice quality in a Japanese feminine style,” *Language in Society*, vol. 44, no. 1, pp. 1–34, 2015.

[13] Stylianou, Y., Cappe, O., and Moulines. E.: “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1988.

[14] Toda, T., Black, A. W., and Tokuda, K.: “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[15] Kobayashi, K., Takamichi, S., Nakamura, S., and Toda, T.: “The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016,” *Proc. of INTERSPEECH*, pp. 1667–1671, 2016. 1st place in speaker similarity

[16] Wu, Z., and King, S.: “Improving trajectory modeling for DNN-based speech synthesis by using stacked bottleneck features and minimum trajectory error training,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, 2016.

[17] Saito, Y., Takamichi, S., and Saruwatari, H.: "Voice Conversion Using Input-to-Output Highway Networks," *IEICE Transactions on Information and Systems*, vol. E100-D, no. 8, pp. 1925–1928, 2017.

[18] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T.: "Speech parameter generation algorithms for HMM-based speech synthesis," *in Proc. ICASSP*, pp. 1315-1318, 2000.

[19] Toda, T., Ohtani, Y., and Shikano, K.: "Eigenvoice conversion based on Gaussian mixture model," *Proc. of INTERSPEECH*, pp. 2446–2449, 2006.

[20] Hsu, C. C., Hwang, H. T., Wu, Y. C., Tsao, Y., and Wang, H. M.: "Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks," *Proc of INTERSPEECH*, pp. 3364–3368, 2017.

[21] Nakashika, T., Takiguchi, T., and Minami, Y.: "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[22] お兄ちゃん CD：Cffon (record label), 2006.

[23] 週刊ノースリー部: [http://www.allnightnippon.com/program/no3b/]

[24] Morise, M., Yokomori, F., and Ozawa, K.: "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[25] Morise, M.: "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[26] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: "GroupLens: an open architecture for collaborative filtering of netnews," *In Proceed-*

ings of the 1994 ACM conference on Computer supported cooperative work, pp. 175–186, 1994.

[27] Chierichetti, F., Kumar, R., Pandey, S., and Vassilvitskii, S.: "Finding the Jaccard median," *In Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics*, , pp. 293–311, 2010.

[28] Hinton, G. E., Osindero, S., and Teh, Y. W.: "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[29] Cho, K., Ilin, A., and Raiko, T.: "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines," *Artificial Neural Networks and Machine LearningICANN 2011*, pp. 10–17, 2011.

[30] Takamichi, S., Tomoki T., Black, A. W., Neubig, G., Sakti, S., and Nakamura, S.: "Post-filters to Modify the Modulation Spectrum for Statistical Parametric Speech Synthesis, " *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

[31] Sedhain, S., Menon, A. K., Sanner, S., and Xie, L.: "AutoRec: Autoencoders Meet Collaborative Filtering," *in Proceedings of the 24th International Conference on World Wide Web*, pp. 111–112, 2015.

[32] Salakhutdinov, R., Mnih, A., and Hinton, G.: "Restricted Boltzmann machines for collaborative filtering," *in Proceedings of the 24th International Conference on Machine Learning*, pp. 791–798, 2007.

[33] 四方田 犬彦: "「かわいい」論," 筑摩書房，2006.

[34] 宇治川 正人："「かわいい」の原因系と結果系の分類," 日本感性工学会論文誌，vol. 15, no. 1, pp. 39–46, 2016.

[35] 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," 日本音響学会誌，vol. 57, no. 5, pp. 337–344, 2001.

付 録A　提案法におけるパラメータの最適化

5.4.2 節で述べた学習では，対数尤度 $\mathcal{L} = \log \prod_n^N p(\boldsymbol{v}|\boldsymbol{\theta})$ を各パラメータ $\boldsymbol{\theta} = \{\overline{\boldsymbol{W}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, \boldsymbol{\sigma}^2, \{\boldsymbol{A}_r\}_{r=1}^R, \boldsymbol{B}, \boldsymbol{C}, \{\boldsymbol{H}_k\}_k, \boldsymbol{B'}, \boldsymbol{C'}\}$ で偏微分した $\frac{\partial \mathcal{L}}{\theta}$（勾配）によって，以下のような最急降下法で各パラメータ $\theta \in \boldsymbol{\theta}$ を最適化する．

$$\theta^{(e+1)} = \theta^{(e)} + \alpha \frac{\partial \mathcal{L}}{\partial \theta} - \beta(\theta^{(e)} - \theta^{(e-1)}) \tag{A.1}$$

ただし，$\theta^{(e)}$ は学習エポックが e 回目時におけるパラメータであり，α, β はそれぞれ学習係数とモーメンタム係数を表す．また，勾配は以下の通りである．

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{H}_k} = < \boldsymbol{v'h}^\top \overline{\boldsymbol{W}}^\top \boldsymbol{A} d_k^\top >_{\text{data}} - < \boldsymbol{v'h}^\top \overline{\boldsymbol{W}}^\top \boldsymbol{A} d_k^\top >_{\text{model}} \tag{A.2}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{B'}} = < \boldsymbol{v'd}^\top >_{\text{data}} - < \boldsymbol{v'd}^\top >_{\text{model}} \tag{A.3}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{C}'_k} = < \boldsymbol{hd}^\top >_{\text{data}} - < \boldsymbol{hd}^\top >_{\text{model}} \tag{A.4}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}_r} = < \boldsymbol{H}^\top \boldsymbol{v'h}^\top \overline{\boldsymbol{W} s_r}^\top >_{\text{data}} - < \boldsymbol{H}^\top \boldsymbol{v'h}^\top \overline{\boldsymbol{W} s_r}^\top >_{\text{model}} \tag{A.5}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{B}} = < \boldsymbol{v's}^\top >_{\text{data}} - < \boldsymbol{v's}^\top >_{\text{model}} \tag{A.6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{C}} = < \boldsymbol{hs}^\top >_{\text{data}} - < \boldsymbol{hs}^\top >_{\text{model}} \tag{A.7}$$

ただし，$\boldsymbol{v'} = \frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2}$ であり，$\overline{\boldsymbol{W}}, \overline{\boldsymbol{b}}, \overline{\boldsymbol{c}}, \boldsymbol{\sigma}^2$ における勾配は文献 [21] と同様である．

付 録B　統計的声質変換による印象の再現性

B.1　はじめに

本章では，統計的声質変換による変換音声の「かわいさ」とそれに関連すると考えられる印象 [33, 34] の変化について調査する．

B.2　実験条件

B.2.1　実験データ

かわいらしい声をもつ代表例として，日本人アイドルである小嶋陽菜（プロダクション尾木）を目標話者に選んだ．声質変換の元話者は 20~21 歳の女子大学生 4 名（$\{s_1,\ s_2,\ s_3,\ s_4\}$）とし，その音声データを使用する．目標話者の音声はラジオ番組 “週刊ノースリー部”[23] から目標話者のみによる明瞭な発話箇所とした．元話者となる 4 名の女子大生には，目標話者の音声と同じフレーズを発話させた．その際に，目標話者の音声表現を模倣しないよう，音声に対応する書き起こしテキストのみを元話者に提示し，元話者自らの音声表現で発話するように指導した．また，ラジオ番組内での他出演者との自然な対話を出来る限り再現するために別途に 1 名の女子大生話者（補助話者）を用意し，元話者は補助話者との対話を行う形で行った．ただし，補助話者にも対話用の書き起こしテキストを提示し，補助話者は自らの音声表現で発話するものとする．

使用する統計的声質変換法として，GMM に基づく最尤系列変換法 [14] を用いた．分析合成は WORLD (D4C edition) [24, 25] で行い，スペクトル特徴量は WORLD から得られるスペクトル包絡をモデル化した 24 次のメルケプストラム及びその

表 B.1: 実験で使用する印象語の一覧

	$1 \Longleftrightarrow 7$	
R1	低い	高い
R2	幼い	大人っぽい
R3	掠れた	澄んだ
R4	暗い	明るい
R5	柔らかい	硬い
R6	滑舌の悪い	はきはきした
R7	たどたどしい	流暢な
R8	弱い	強い

R9	鼻音	通った
R10	かわいくない	かわいい
R11	男性的	女性的
R12	落ち着きのない	落ち着いた
R13	太い	細い
R14	張りのある	張りのない
R15	親近感のある	親近感のない
R16	無邪気な	悪意のある

動的特徴量とし，シフト長は 5 ms，サンプリング周波数は 16 kHz とした．また，GMM の混合数は 128 である．学習に用いる音声は 120 文，評価に用いる音声は 30 文とする．このとき，評価音声には，目標話者特有のフレーズ（自己紹介や友人の名前などを挙げているもの）は含まれないようにした．

B.2.2　聴取実験

s_1，s_2，s_3，s_4 の 4 名の話者を知らない 20 人（男女 10 人ずつ，18 ～ 22 歳）に対して 2 種類の聴取実験を行った．元話者 4 名の音声，その変換後の音声（話者 s_i の変換後の音声を，便宜上「話者 s_i' の音声」と呼ぶ），目標話者の音声の各々に対し，印象語の評価を 7 段階で行う．このとき評価者の負担を考慮し，各評価者共に 81 個の音声のみ評価させた（各音声刺激は 6 人によって評価される）．「かわいさ」は “幼い”，“弱い”，“明るい”，“女性的”，“落ち着きない”，“親近感のある”，“無邪気な” などの印象と関係が深い [33, 34]．これらを踏まえた上で，印象語は声質の印象語に関する研究 [35]，及び「かわいさ」に着目した研究 [33, 34] を参考に選んだ（Table B.1）．また，同評価者に変換音声の個人性の再現度を DMOS 評価に従い 5 段階で評価させた．ここでも評価者の負担を考慮し，各評価者共に 56 個の音声のみ評価させた（各音声刺激は 10 人によって評価される）．

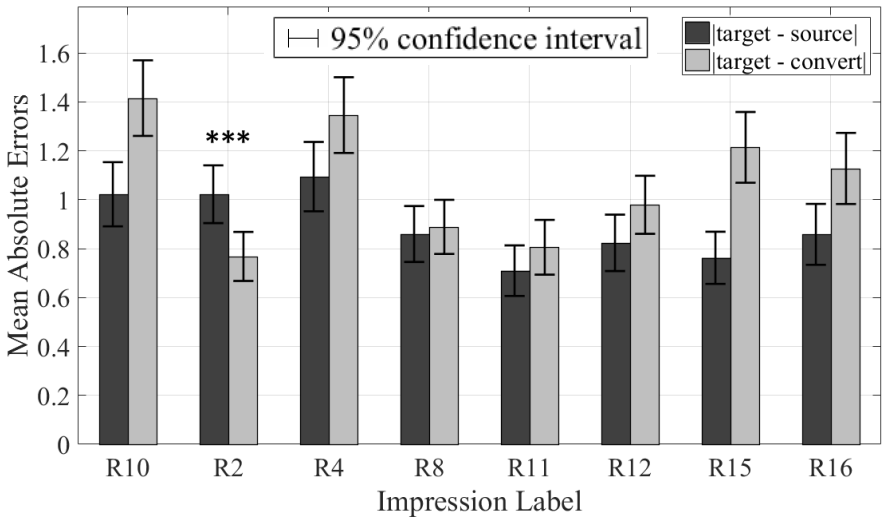

図 B.1: 印象評価の結果（横軸は各印象語に対応する．$***$, $**$, $*$ は検定でそれぞれ $p < 0.01$, $p < 0.05$, $p < 0.1$ を表す.）

B.2.3　分析方法

今回は被験者の負担を考慮し，各被験者は一部の音声刺激にしか評価を行っていないため，各音声刺激は被験者 6 人分の評価値を付与されている．そこで，各音声刺激，各印象語毎に 6 人の評価者による評価値の平均値を算出し，その音声刺激の各印象語に対する評価とする．

s_i と目標話話者，s'_i の各印象語の評価値に関し，s'_i と目標話者の差の絶対値が，s_i と目標話者のそれより有意に小さければ，再現性が高いと言える．今回，標本の大きさが小さく，また正規性や等分散性の保証ができないため，Brunner-munzel の検定（有意水準 5%の片側検定）を用いる．

B.3　実験結果・考察

印象評価の結果を図 B.1 に記す．図 B.1 は，各印象語に対する評価値において，変換元話者 s_i と目標話者の差の絶対値及び変換後の話者 s'_i と目標話者の差の絶対値を平均したものである（エラーバーは 95% 信頼区間）．R2（大人っぽい）で目標話者に有意に近付いていることがわかる（$p < 0.01$, 統計量 $W = -3.18$）.

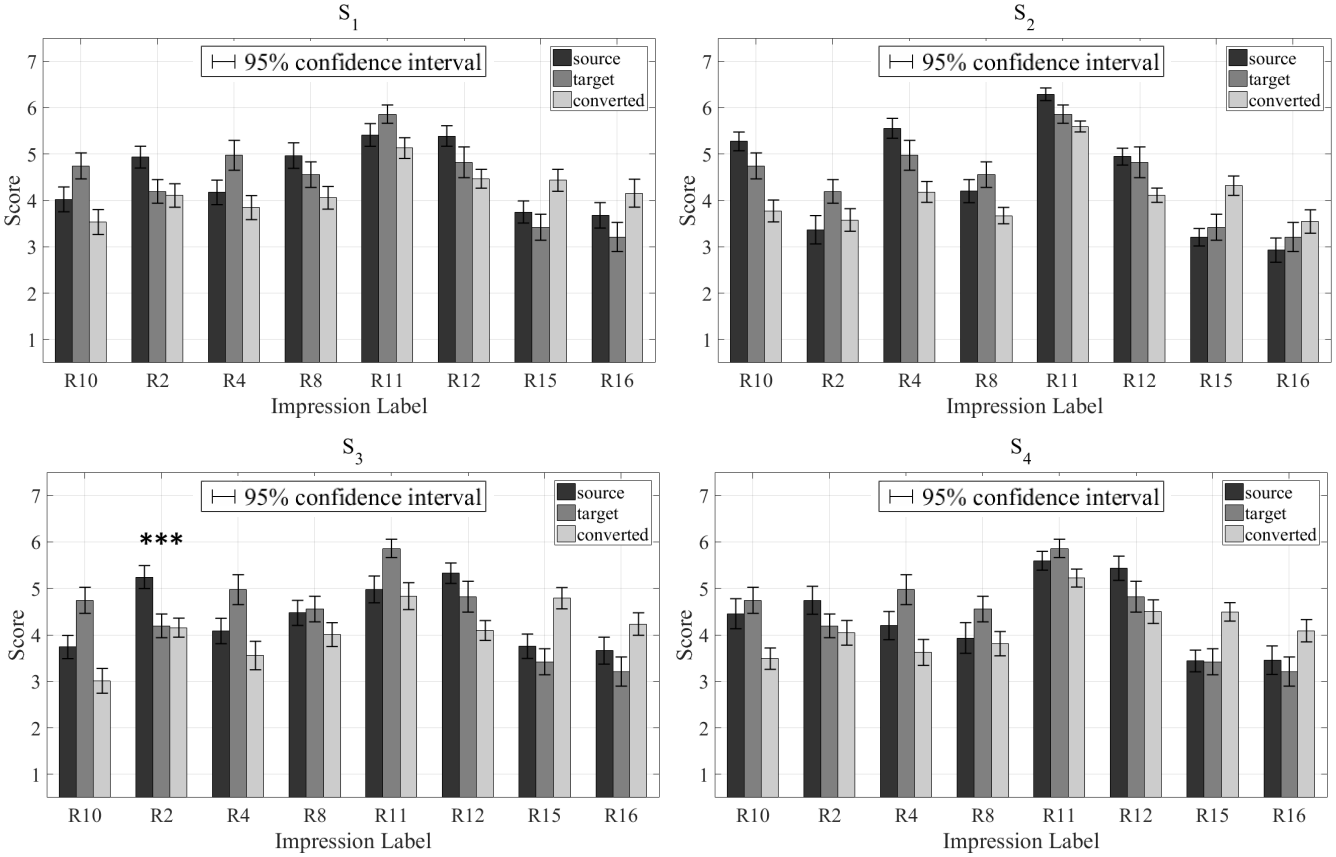

図 B.2: 話者毎の印象評価の結果（横軸は各印象語に対応する．$***$, $**$, $*$ は検定でそれぞれ $p < 0.01$, $p < 0.05$, $p < 0.1$ を表す.）

次に，話者毎の印象評価の結果を図 B.2 に記す．図中の Score は，s_i，目標話者及び s'_i それぞれの印象評価値の平均である．s_3 は R2（大人っぽい）で目標話者に有意に近付いていることがわかる（$p < 0.01$, 統計量 $W = -3.43$）．一方で，R4（明るい），R16（悪意のある），さらに R10（かわいい）は，s_2 以外では変換話者の印象が目標話者のものから遠ざかるように変化している．s_2 に関しては印象変化の方向は目標のものへ近付く方向へ変化しているが，過剰に変化している．R15（親近感のない）では，s_1 と s_3 の変換話者の印象は目標から遠ざかるように変化しており，s_2 と s_4 の場合は目標に近付くように変化しているが，過剰に変化している．声質変換では変換音声は自然性の劣化が生じてしまい，今回の手法でも例外ではない．被験者にとって，今回のような自然性の劣化が含まれている合成音声

（低品質な合成音声）に対して良い印象を持つのが難しいと考えられるため，この劣化が “暗い”，“親近感のない”，“悪意のある”，さらに “かわいくない” という方向へ印象が変化（s_2 の場合は過剰に変化）している原因として挙げられる．

以上より「かわいさ」と関係のある印象のうち，“幼い” のみは再現できている傾向にあるが，自然性劣化の影響を強く受けてしまう印象に関しては,「かわいさ」から離れる方向に変化するため，再現ができなかったと考えられる．

また，全 16 種類の印象全てに対する評価結果を図 B.3 と図 B.4 に記す.

B.4 本章のまとめ

本章では統計的声質変換技術を用いて生成した音声の聴感上の「かわいさ」に着目し,「かわいさ」及びそれに従属すると考えられる印象がどう変化するかを調査した.「幼い」に関しては再現出来ていた一方で，自然性の劣化による影響が大きいと考えられる印象に関しては正しい評価を行うことが難しく，目標話者に対する値に寄らず過剰に印象が変化することがあった.

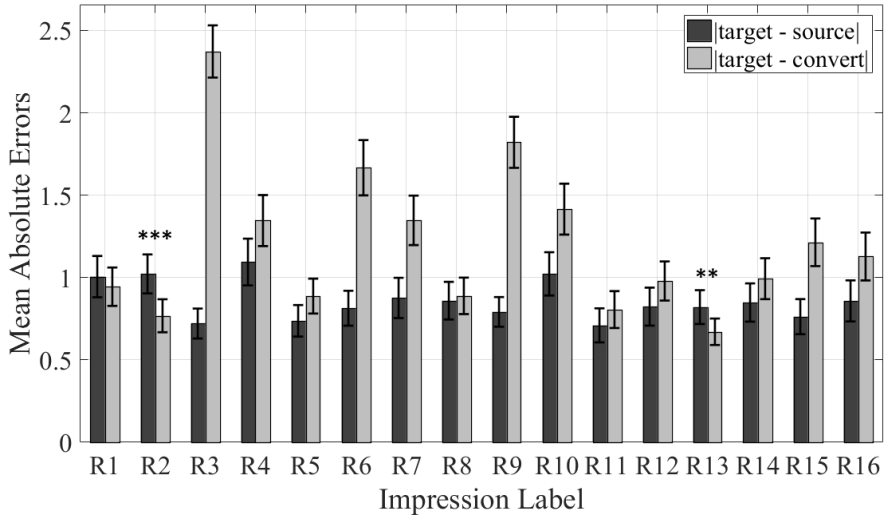

図 B.3: 印象評価の結果（横軸は各印象語に対応する．$***$, $**$, $*$ は検定でそれぞれ $p < 0.01$, $p < 0.05$, $p < 0.1$ を表す.）

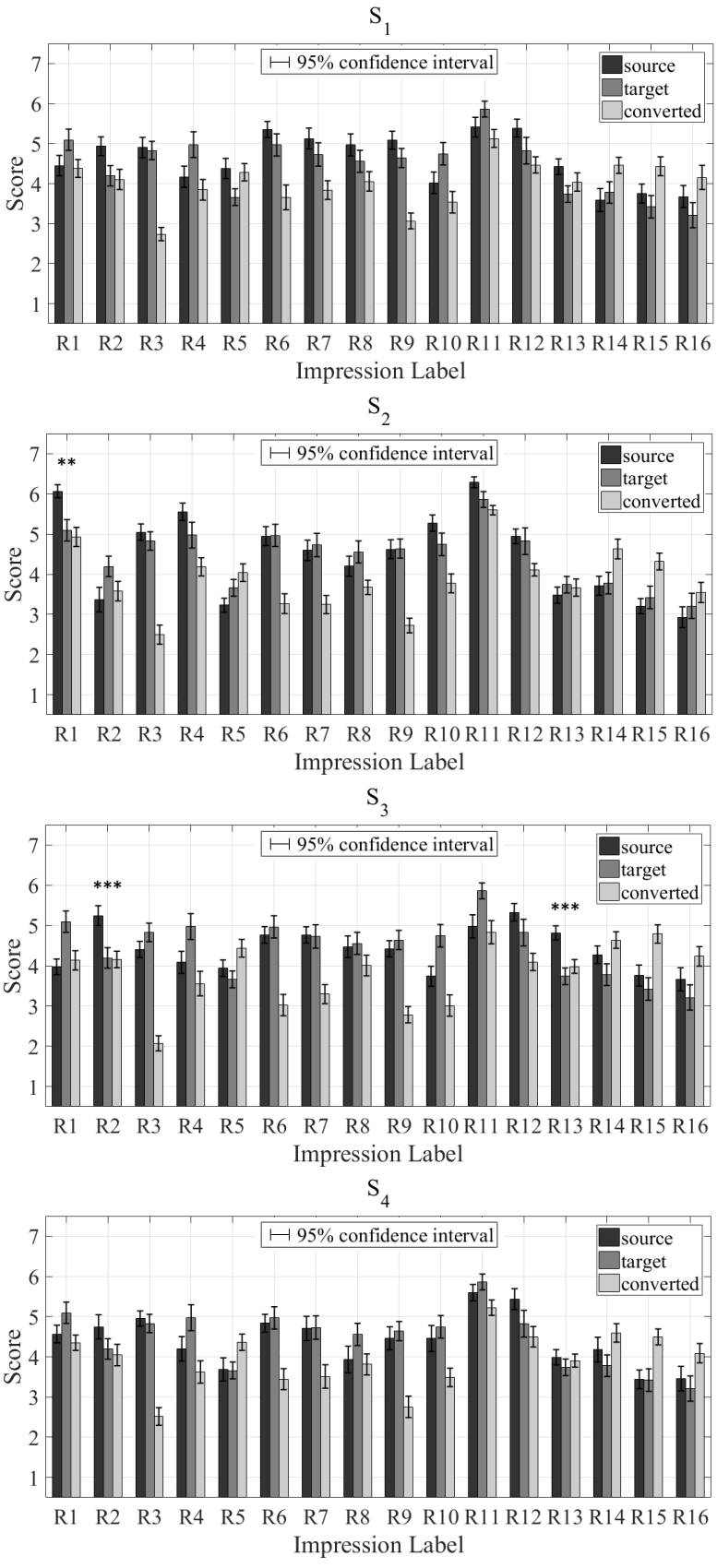

図 B.4: 話者毎の印象評価の結果（横軸は各印象語に対応する．$***$, $**$, $*$ は検定でそれぞれ $p < 0.01$, $p < 0.05$, $p < 0.1$ を表す.）

発表リスト

国際会議

1. Ohno, R., Morise, M., and Kitahara, T.: "Relationship between perception of cuteness in female voices and their durations," *Lecture Notes in Computer Science (in Proc. of SPECOM)*, LNCS, 10458, pp. 642–650, 2017.
2. Ohno, R., Morise, M., and Kitahara, T.: "The relationship between perception of cuteness and duration of voices," *Joint Meeting : Acoustical Society of America and Acoustical Society of Japan.*, vol. 140, no. 4, p. 3399, 2016.

研究会

1. 大野涼平，森勢将雅，北原鉄朗:"音声における「かわいらしさ」の知覚と聴取時間の関係性の検討," 情報処理学会音楽情報科学研究会, vol. 2016-MUS-111, no. 50, pp. 1–5, 2016. (学生奨励賞)

全国大会（主著）

1. 大野涼平，高道慎之介，森勢将雅，北原鉄朗："話者適応型 RBM を用いたユーザが所望するかわいい音声への声質変換," 日本音響学会 2018 年春季研究発表会，2018. （発表予定）
2. 大野涼平，高道慎之介，森勢将雅，北原鉄朗："音声の「かわいさ」における主観的傾向の一分析," 日本音響学会 2017 年秋季研究発表会, pp. 401–402, 2017.
3. 大野涼平，高道慎之介，森勢将雅，北原鉄朗："統計的声質変換における印象評価の調査," 日本音響学会 2017 年春季研究発表会, pp. 371–372, 2017.

全国大会（共著）

1. 松下禎希，大野涼平，北原鉄朗："雑音が記憶や作業に与える影響に関する一調査，" 日本音響学会 2018 年春季研究発表会，2018. （発表予定）
2. 長谷川翔太，大野涼平，北原鉄朗："男性両声類の女声らしさに関わる特徴量の分析," 日本音響学会 2017 年春季研究発表会，pp. 371–372, 2017.

謝　　辞

本論文を作成するにあたり，本学の北原鉄朗准教授には指導教員として，丁寧で熱心な指導を賜りました．自由なテーマで研究に取り組ませていただき，研究の面白さを教えてくださいました．心より感謝申し上げます．

副査を担当してくださった本学の斎藤明教授と古市茂教授には，細かな添削をしていただき，本質的な鋭いご指摘やご意見をいただきました．また，本学の宮田章裕准教授には厳しくも温かいコメントをいただき，貴重な議論をしていただきました．心から感謝申し上げます．

東京大学の高道慎之介特任助教には，音声信号処理や声質変換に関する多くの助言をいただきました．原稿，資料作成時には大変丁寧な添削やご指導をしていただきました．また，いつでも親身に相談に乗ってくださり，多くの指導のお陰で，この分野の研究の面白さを知ることができました．心より厚く感謝致します．

山梨大学の森勢将雅准教授には，筆者が本学学部 4 年時から技術的なアドバイスだけでなく，研究への姿勢など，多岐に渡るアドバイスをいただきました．原稿，資料作成時には細かな指導をしていただき，また，行き詰まっているときには温かい言葉で励ましていただきました．心より深く感謝致します．

卒業生の鈴木潤一氏と栗原拓也氏とは，お二人が在学中に研究に限らず多くの助言をいただきました．また，プライベートでも仲良くしていただき，居心地の良い環境にしてくださりました．心より感謝申し上げます．

同期の棚橋徹氏とは時には励まし合い，時には熱い議論を交わし，切磋琢磨してきました．また，プライベートでも濃い時間を過ごし，2 年間苦楽を共に過ごした棚橋氏に心より厚く感謝致します．

また，居心地の良い環境にしてくださった本研究室の後輩の皆様にも，心より深く感謝致します．

多くの方々のお力添えがあり，本論文をまとめることができました．心からの

感謝の意を表します.