Vol.2021-SLP-136 No.31
2021/3/4

リアルタイムDNN音声変換フィードバックによるキャラクタ性の獲得手法

倉田 将希[1,a)]　高道 慎之介[1,b)]　佐伯 高明[1]　荒川 陸[1]　齋藤 佑樹[1]　樋口 啓太[1]　猿渡 洋[1,c)]

概要：本稿では，音声変換ユーザに目標話者のキャラクタ性を獲得して発話させるためのシステムを提案する．深層学習に基づくリアルタイム音声変換は，人間の発声器官の物理制約を超えて，ユーザの音声から所望のキャタクタ性を持つ音声への高精度な変換を可能にしつつある．しかしながら，音声のパラ言語情報（抑揚・強勢など）の変換は未だ困難であり，ユーザの音声のパラ言語情報が変換音声に直接的に反映されてしまう．また，通常の発話において，人間は自己聴取音の聴取との相互作用により自らの言語情報・パラ言語情報を制御するが，リアルタイム音声変換を用いた発話において，そのような相互作用をもたらす機構は存在しない．そこで本稿では，変換音声をユーザにリアルタイムにフィードバックする自己聴取音制御システムにより，変換音声に所望のキャラクタ性を付与するようユーザを発話変容させるシステムを提案する．実験的評価では，一人称視点（音声変換ユーザ視点）と三人称視点においてシステムおよび変換音声を評価し，(1) 演技経験の少ないユーザに対してシステムの有用性が高いこと，(2)F_0 を目標キャラクタに近づけるだけで十分な発話変容効果がみられることを示す．

Masaki Kurata[1,a)]　Shinnosuke Takamichi[1,b)]　Takaaki Saeki[1]　Riku Arakawa[1]　Yuki Saito[1]
Keita Higuchi[1]　Hiroshi Saruwatari[1,c)]

1. はじめに

音声変換（Voice Conversion: VC）とは，入力話者音声の持つ言語情報を保持したまま，非言語情報のみを変換する技術の総称である．別の人物の発話へと変換する話者変換 [1] をはじめ，外国語教育支援 [2] や吃音治療 [3] での利用など，音声変換の応用は多岐に渡る．計算機が発達した近年では，統計的なアプローチを用いた研究 [4], [5], [6] も広く行われ，音声特徴量の変換規則を統計的に学習することで，高精度な変換が実現されている．Deep neural network（DNN）に基づく非線形な特徴量変換を用いた手法 [5], [6] はその代表例であり，Gaussian mixture model（GMM）を用いた区分線形変換の手法 [4] に比べて，高音質な音声変換が可能である．また，深層学習に基づく音声変換は，リアルタイムに動作するアルゴリズム [7], [8] も確立されており，このリアルタイム音声変換のさらなる発展により，人間の発声器官の物理的制約を超えた新しいコミュニケーション社会の実現も期待される．

しかしながら，現在の音声変換はパラ言語情報（抑揚・強勢など）の変換における問題を孕む．パラ言語情報は発話意図や感情を伝える音声的特徴があり [9], [10]，発話者の個人性が強く表れる．しかし，典型的な音声変換におけるパラ言語情報の変換は単純な音声特徴量変換規則に従う [4], [11] ため，音声変換ユーザによる入力音声に含まれるパラ言語情報が，変換音声に直接的に反映されてしまう．故に，特徴的なパラ言語情報を持つ目標話者（以降，目標キャラクタ）に音声変換する場合，その変換音声のキャラクタ再現度は非常に低いものとなる．この解決方法として複雑な変換モデルを使用する方法が模索されている [12], [13] が，その効果は限定的である．さらに，変換のリアルタイム性とモデルの複雑性はしばしばトレードオフの関係にあるため，リアルタイム変換を前提としたパラ言語情報の変換は，さらに困難なものとなる．現在の音声変換に対する別の問題点として，speech chain 構造 [14] の欠落がある．speech chain とは，音声変換を介さない通常の音声生成において，発話者が自らの発話音声を聴取した結果に基づいて音声生成器官を制御する一連の処理の連鎖のことである

[1] 東京大学, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] 98m7-disney-0n-classic@g.ecc.u-tokyo.ac.jp
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[c)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

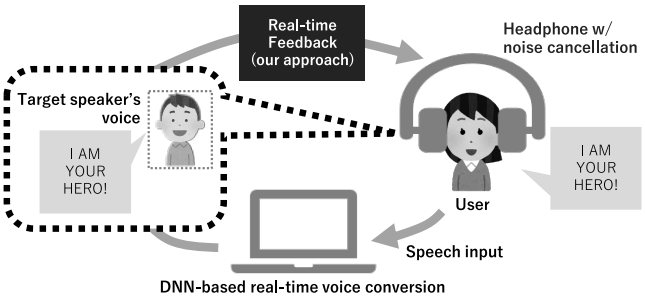

図 1 提案システムの概要

が，これまでのリアルタイム音声変換の研究では，第三者視点の変換品質のみが評価されており，聴覚フィードバック及び相互作用が考慮されてこなかった．

本稿では，これらの問題を改善するため，変換音声をユーザにリアルタイムにフィードバックし，発話変容を促すシステムを提案する．自己聴取音のうち空気伝導音をリアルタイム DNN 音声変換により変換し，発話者にリアルタイムにフィードバックする．提案システムにより音声変換を要素に含む speech chain が構成されるため，変換音声に所望のキャラクタ性を付与するようユーザを発話変容させる効果が期待される．実験的評価では，アニメキャラクタの模倣をタスクとし，提案システム及び変換音声に対し，発話者自身による一人称評価と第三者による三人称評価を実施する．実験的評価の結果，演技経験の少ない話者に対する有用性が高い傾向にあること，F_0 を目標話者に近づけるだけで十分な発話変容効果がみられることを示す．

2. リアルタイム DNN 音声変換 [7]

Arakawa ら [7] は特徴量解析部・特徴量変換部・波形合成部から構成されるリアルタイム DNN 音声変換手法を提案し，各処理を再帰的かつ低遅延に行うことでアルゴリズム遅延 50 ms を実現している．また，人工的にデータを水増しするデータ拡張により，合成音声の音質を向上できることを報告している．DNN に基づくスペクトル変換により非言語情報は高精度に変換される一方で，パラ言語情報はパワーと F_0 の線形変換のような限定的な変換に留まる．

3. 提案システム

3.1 リアルタイム音声変換を含む聴覚フィードバック

本研究で提案する，リアルタイム音声変換を含む聴覚フィードバックシステムの枠組みを図 1 に示す．計算機に入力されたユーザの音声は，2 節の手法により逐次的に変換され，音声変換波形が合成される．この変換音声は，ユーザの装着するヘッドホンに出力される．以上のシステム構成により，リアルタイム音声変換を含む speech chain が構成される．

3.2 外音抑圧による空気伝導音の抑圧

通常の音声生成において発話者にフィードバックされる自己聴取音声は，その伝達経路の違いから空気伝導音声と体内伝導音声に分けられる．3.1 節の手法は，この空気伝導音声をリアルタイム音声変換により変換するものと見做される．しかしながら，発話者本来の空気伝導音声の伝達経路（すなわち，口唇から外耳道）は依然存在するため，本来の空気伝導音声と変換音声の両方が，変換処理だけの時間的ズレを以って発話者にフィードバックされることになる．予備調査では，この条件下の発話は吃音症状を生み，目標話者のキャラクタ性を付与するようなユーザの発話変容を妨げることが確認された．そのため提案システムでは，密閉型ヘッドホンと active noise cancellation により，本来の空気伝導音声を抑圧する．

4. 実験的評価

4.1 実験条件

提案するシステムでは 2 節で示したリアルタイム DNN 音声変換手法を使用し，パラメータも Arakawa らの論文 [7] に従った．サンプリング周波数は 16 kHz，変換システム全体の遅延は約 190 ms であった．この値は，計算機の音声入出力に係る遅延（約 140 ms）と変換に係る遅延（約 50 ms）の和である．外音抑圧と聴覚フィードバックには Sony 製ヘッドホン WH-1000XM4 [15] を使用した．このヘッドホンは Bluetooth による無線接続機能を有するが，Bluetooth 接続による処理遅延を除外するため本実験では有線接続を用いた．以降の節に示す一人称実験（事前アンケート及び実験 1）は，東京大学無響室[*1]において実験参加者単独で実施した．変換先の目標キャラクタは，日本語アニメ作品に登場する男性日本語母語キャラクタ 1 名とした．変換音声ユーザとなる被験者は，表 1 に示す 14 名（20 代：13 名，30 代：1 名）である．この被験者は，縁故法により募集し，演技経験の有無については，被験者本人の申告から得た．リアルタイム DNN 音声変換の学習用音声を作成するため，各被験者の音声を事前に収録した．この事前収録は以降の節に示す実験と別日に行い，ATR 音素バランス文 [16]A01 から B45 の 95 文を各被験者に読み上げさせた．この事前収録に関する予備調査として，学習用音声と変換時音声の発話スタイルを一致させることで変換音声の品質向上が認められた．そのため，事前収録において，目標キャラクタと同程度の F_0 レンジを持つ音声を見本として提示し，被験者に対し，その F_0 レンジにできる限り合わせて発話するよう指示した．ここで，目標キャラクタの音声を提示しなかったのは，事前収録において被験者が

[*1] ただし，COVID-19 対策として空気循環用サーキュレータを稼働した状態で実験を実施したため，無響状態ではない．収録音声にサーキュレータ稼働音の影響が認められたものの，音声に対する稼働音のパワーは十分に小さいため，本実験においてサーキュレータの影響を無視する．

表 1 一人称主観評価実験の被験者分布 (性別・演技経験別)

	演技経験あり	演技経験なし	計
男性	2	6	8
女性	3	3	6
計	5	9	14

目標キャラクタのキャラクタ性を獲得してしまう学習効果を避けるためである．各被験者と目標キャラクタの学習用音声から，被験者毎の音声変換システムを構築した．以降の節に示す三人称実験（実験 2）の参加者（以降，第三者評価者）は，縁故法により募集した 11 名である．なお，被験者と第三者評価者は重複しない．三人称実験に使用する音声変換には，非リアルタイム DNN 音声変換手法 [17] を採用した．これは，できる限り高い音声変換品質を達成することで，第三者評価者が変換音声のキャラクタ性を適切に評価できるようにするためである．三人称実験は，第三者評価者が単独で参加し，静音環境のもとヘッドホンで音声サンプルを聴いて行った．第三者評価者による評価数は，以降の節に示すフィードバック対（図 4）について 44，性別及び演技経験の対（図 5, 6）について 66 である．

4.2 実験手順

以下の一人称実験（事前アンケート，実験 1）を被験者に，三人称実験（実験 2）を第三者評価者に対し実施する．

事前アンケート：キャラクタへの親密度の測定

被験者が目標キャラクタをどの程度知っているかに関するアンケートを行う．質問項目は

- 目標キャラクタの登場するアニメをどの程度観た経験があるか
- 目標キャラクタをどの程度知っているか

であり，それぞれに対し “あまり観ていない (知らない)”，“ある程度観ている (知っている)”，“非常に観ている (知っている)” の主観的キャラクタ親密度に関する 3 択を被験者に回答させる．また，客観的なキャラクタ親密度を図るために，目標キャラクタの登場するアニメの公式ウェブサイトの文面から目標キャラクタに関する 8 項目を作成し，各項目の内容を知っているか否かを被験者に回答させる．8 項目中 “知っている” と回答した項目数を客観的なキャラクタ親密度として求め，0–4 項目 (“あまり知らない”)，5–6 項目 (“ある程度知っている”)，7–8 項目 (“非常に知っている”) の 3 段階でランク分けする．

実験１：フィードバック実験と一人称評価アンケート

まず，目標キャラクタのキャラクタ性を被験者間で統一するため，被験者に対し目標キャラクタの字幕・音声つき映像を提示する．これは，アニメにおける目標キャラクタの登場シーンと，目標キャラクタの学習用音声からなる 5 分程度の映像である[*2]．なお，この映像に登場するセリフは，以降に示すフィードバック実験で用いるセリフと重複しない．

次に，被験者に発話セリフのテキストを提示し，以下に示す 4 種類のフィードバック条件下で目標キャラクタを模倣して発話させる．

no FB：音声変換・音声フィードバック (Feedback: FB) を行わず，被験者は提示テキストのみを用いて模倣する．ヘッドホンは装着する．

non-chara FB：目標キャラクタではない参照話者の音声に変換した音声をリアルタイムにフィードバックする．

chara FB：目標キャラクタの音声に変換した音声をリアルタイムにフィードバックする．

chara FF (reference)：音声変換・音声フィードバックを行わないが，被験者は提示テキストに加え，目標キャラクタの実発話音声を聴取 (Feed-forward: FF) できる．ヘッドホンは装着する．

non-chara FB において，この参照話者と目標キャラクタの F0 レンジは同程度であるため，non-chara FB と chara FB の変換音声の間で異なる音声特徴量は，音色を付与するスペクトル特徴量のみである．各被験者で各フィードバック条件をランダム順に実施する．発話セリフは，アニメにおいて目標キャラクタの発話する高模倣親密度[*3]セリフ 20 文と，ATR 音素バランス文から抽出した低模倣親密度セリフ 20 文（学習用音声とは重複しない）から成る計 40 文である．各被験者と各フィードバック条件において，ランダムに抽出した 10 文 (高模倣親密度セリフ 5 文と低模倣親密度セリフ 5 文) のテキストを被験者に提示し，キャラクタを模倣させる．この際，後述 r する実験 2 のために，被験者の発話音声を録音する．

各フィードバック条件における実験が終了後，以下の 6 項目に関する実験後アンケートを実施し，各項目において 5 段階 MOS テスト形式で絶対評価させる．ただし，最後の 4 項目はフィードバックあり 2 手法のみの実施である．

Immersion：セリフの発話に没入できたか．

Similarity：目標キャラクタを模倣して発話できたか．

Delay：フィードバックの遅延が気になったか．

Quality：変換音声の音質が気になったか．

Usability：システムが模倣に役立ったか．

Intention：システムを今後も使いたいか．

以上 6 項目のほか，自由記述式アンケートを実施する．

実験２：三人称視点による音声評価

実験 1 の各フィードバック条件における音声を第三者評価者により評価させる．実験 1 において録音した被験者の

[*2] 紙幅の都合上，詳細を省略するが，「映像の提示は模倣に役立ったか」の 5 段階評価（1：役に立たなかった〜5：役に立った）に対し，7 割以上の被験者が 4 以上を回答した．

[*3] “被験者による発話模倣のし易さ” の度合いを模倣親密度と定義し，その高低を与えた．紙幅の都合上，詳細な説明を省略するが，実験後アンケートにおいて各発話セリフの模倣し易さを調査したところ，その上位を高模倣親密度セリフが占めた．

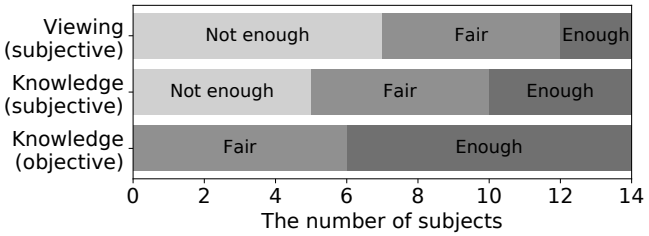

図 2 主観的キャラクタ親密度（上・中央）と客観的キャラクタ親密度（下）

音声を，目標キャラクタの声に変換し，以下の 2 項目に関して別々に評価させる．

Speech naturalness：変換音声はどの程度自然か

Character similarity：変換音声はどの程度目標キャラクタ性を再現できているか．

各項目を，それぞれプリファレンス AB テスト・XAB テストにより評価する．異なるフィードバック条件対 (ただし，実験 1 で reference として設定した chara FF 条件は除く)・性別対・演技経験対それぞれの変換音声を，第三者評価者にランダム順に提示し，自然性もしくはキャラクタ再現度の高い方を別々に選択させる．対提示する音声を受聴する回数と順番に制限は設けない．XAB テストを実施する際，目標キャラクタのキャラクタ性を第三者評価者間で統一するため，実験 1 と同様の映像を第三者評価者に事前に提示する．このとき第三者評価者は，映像を視聴した後に 2 条件の音声を受聴し，キャラクタ再現度を評価する．映像を視聴する回数の制限は設けないが，評価実験開始後に再生することは禁止する．前述の通り，この映像中のセリフと変換音声のセリフは重複しない．

4.3 実験結果

事前アンケート結果：キャラクタへの親密度

キャラクタへの親密度に関するアンケートの結果を図 2 に示す．主観的には，アニメを観た経験やキャラクタについて知っている度合を高く評価した被験者は少ないことがわかる．客観的なキャラクタ親密度では，被験者は一定水準以上，目標キャラクタについて知っていると判断される．以上より，本実験の被験者は，主観的キャラクタ親密度のばらつきはあるものの，一定以上の客観的キャラクタ親密度を持つ集団といえる．

結果 1：一人称評価アンケート

前述の 6 項目に関する 5 段階評価に対して，被験者の演技経験の有無 (Performance)，フィードバック条件 (Feedback)，及び両者の交互作用 (Interaction) の各要因が影響するか否かを，二元配置分散分析の混合計画 (two-way ANOVA mixture) により検定し，表 2 に結果 (p 値) を示す[*4]．検定の結果, Immersion の評価には演技経験とフィードバック条件がそれぞれ影響し ($p < 0.05$)，Similarity の評価にはフィードバック条件のみが影響し，その他の項目はいずれの要因も影響するとはいえない ($p > 0.05$) ことがわかった．ただし，Usability は演技未経験者の評価が高い傾向にあった ($p = 0.185$)．各要因の交互作用が影響する項目はなかった．

[*4] いずれの検定においても自由度は df1=1, df2=12 であった．

表 2 ANOVA による検定結果 (p 値)

項目	Performance	Feedback	Interaction
Immersion	0.044	0.006	0.646
Similarity	0.137	0.000	0.397
Delay	0.883	0.306	0.117
Quality	0.803	0.814	0.752
Usability	0.185	0.721	0.789
Intention	0.538	0.472	0.590

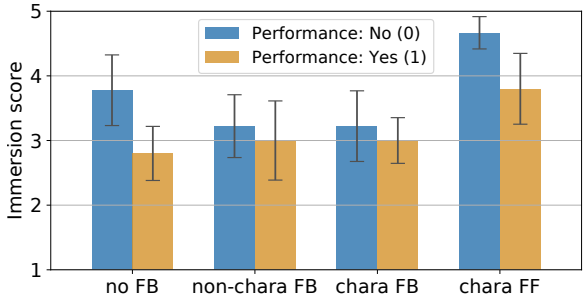

図 3 Immersion に対する評価の平均値と標準偏差 (演技経験別)

Immersion に対する演技経験別の評価結果を図 3 に示す．演技未経験者（Performance: 0）の評価が一貫して高いことがわかる．これと表 2 より，セリフの発話の没入感は演技未経験者の方が高いと回答する傾向があるといえる．

表 2 でフィードバック要因の影響が確認された項目に対して，どの群の組み合わせにおいて有意差があるのかを Benjamini/Hochberg FDR correction 法を用いて検定した結果が表 3 である[*5]．Immersion 及び Similarity の両項目とも，chara FF とその他の 3 手法の間にのみ有意差が確認された ($p < 0.05$)．また，Similarity では no FB と chara FB の評価の間に差がある傾向はみられた ($p \fallingdotseq 0.107$)[*6]．以上により，セリフの発話の没入感と目標キャラクタの模倣度合に対するユーザの評価は，目標キャラクタによる実発話音声を聴取することで向上し，実発話音声がない場合はフィードバックを行うことによる統計的な有意差はないといえる．ただし模範度合は，目標キャラクタの音声に変換してフィードバックする場合に向上する傾向があった．

フィードバックのある 2 手法が表 3 の各項目で no FB に有意差を持たなかった原因の一つには，フィードバック遅延による喋りづらさが考えられる．記述アンケートにおいても半数以上の被験者が指摘しており，より低遅延なシステム構築が求められる．

変換される F_0 レンジが同程度である non-chara FB と

[*5] 本検定手法で p 値は補正されるため，補正後の p 値のみ示した．
[*6] 評価の平均値は chara FB の方が高かった．

表 3　要因内の有意差検定結果 (p 値)

A	B	Immersion	Similarity
no FB	non-chara FB	0.643808	0.470247
no FB	chara FB	0.620900	0.107183
no FB	chara FF	0.003163	0.000380
non-chara FB	chara FB	1.000000	0.469043
non-chara FB	chara FF	0.025794	0.004432
chara FB	chara FF	0.002678	0.002758

chara FB の条件間で，いずれの項目においても有意差が確認されなかったことは「ユーザにキャラクタ性を付与するような発話変容を促すには変換音声の F_0 のみを目標話者に近づければよく，スペクトル特徴量を変換することの効果は小さい」(仮説 1) ということを示唆する．これが明確に示されれば，DNN 学習や非線形変換にかかるコストや時間を削減しリアルタイム性を向上することも期待できるが，被験者の中には chara FB 特有のアシスト効果を述べた意見もあった．

演技経験の有無と没入感評価の違いに関連して，記述アンケートでは「模倣しようとイメージするほど自分のイメージに没入するようになるため音声アシストが耳に入りづらくなった」(演技経験者),「no FB は素の自分の声を聴いてしまい役に入り込めなかったがフィードバックありの手法は自分でない声が聴けるため没入しやすかった」「間の取り方を調整することで変換音声を目標キャラクタに近づけられたように思う」(演技未経験者) などが挙げられた．これと図 3 及び表 2 の結果より，目標キャラクタを模倣する際に「演技経験者は音声アシストよりも自らのイメージに没入する一方で，演技未経験者は自らの声よりも音声アシストを活用する形でイメージに没入するため，音声のフィードバックや実発話音声の提示の効果は演技未経験者において高くなる傾向にある」(仮説 2) ことが考えられる．

結果 2：三人称視点による音声評価結果

フィードバック条件対の第三者評価の結果を，ユーザの性別ごとに図 4 に示す．自然性は男女とも no FB, non-chara FB, chara FB の順に高く評価された一方，キャラクタ性再現度は男女間で傾向が異なり，男性では chara FB，女性では no FB が高く評価されている．性別対の評価は図 5 に示す．自然性に一貫性はなく，キャラクタ性再現度は男性被験者が高く評価されている．演技経験対の評価は図 6 に示す．自然性は no FB で有意差がなく，フィードバックありの 2 手法では演技経験者の発話が高く評価されている．一方でキャラクタ性の再現度は，no FB で演技経験者，その他 2 手法で演技未経験者の発話が高く評価された．

音声フィードバックを行った 2 手法の発話に対する自然性評価が低かったのは，前述のフィードバック遅延による吃音症状が原因として考えられ，改善が求められる．

本実験の目標キャラクタ及び参照話者は両者とも男性だが，その F_0 レンジは女性に近い (平均値が 215 Hz 程度).

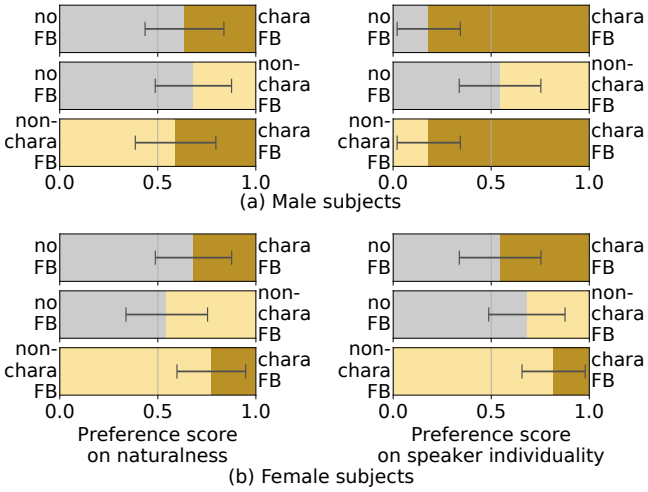

図 4　フィードバック条件対比較の結果 (平均と 95%信頼区間)．灰色: no FB，黄色: non-chara FB，黄土色: chara FB.

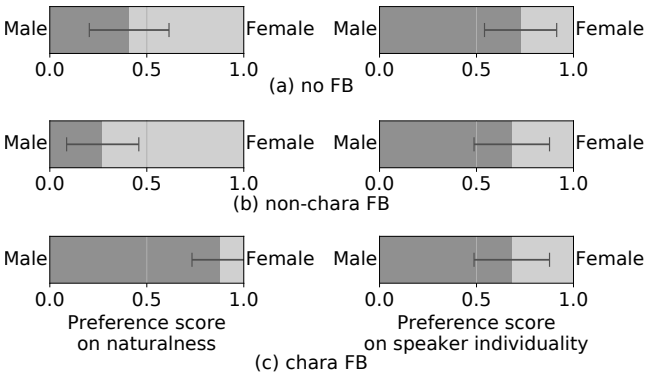

図 5　性別の対比較の結果 (フィードバック条件別，平均と 95%信頼区間)．濃灰色: 男性，薄灰色: 女性.

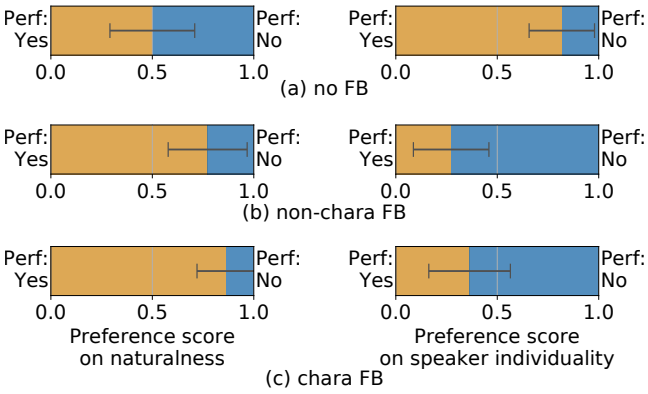

図 6　演技経験の対比較の結果 (フィードバック条件別，平均と 95%信頼区間)．“Perf.” は Performance(演技経験) の略．黄色: 演技経験者，青色: 演技未経験者.

音声変換は通常，F_0 レンジの近い話者間において品質が向上するため[*7]，女性被験者へのフィードバックの効果が期待される．しかしながら，図 5 に示すように逆の結果が得られ，男性被験者へのフィードバックの効果が認められた．これは，高品質な音声変換がキャラクタ性の再現度を

[*7] 実際に，本実験の変換音声を被験者性別間で比較した場合，女性被験者の自然性スコアが高くなる傾向になった.

向上させるとは限らないことを示唆する．男性被験者への効果が得られた理由として，被験者と目標キャラクタの性別の一致が考えられるが，さらなる調査については今後の課題とする．

no FB 条件で演技経験による自然性評価の差がない一方で，キャラクタ性の再現度は演技経験者の方が高く評価されたことは，特定の話者を模倣する行為に対して演技経験者が一定水準の資質を持ち合わせていたことを示唆し，演技経験に対する被験者本人の申告の妥当性を裏付ける．

フィードバックを行った 2 手法において，演技経験者の発話は自然性が高くキャラクタ性の再現度が低いのに対し，演技未経験者の発話には全く逆の傾向がみられたことは，前述の仮説 2 を応用する形で「音声フィードバックを積極的に活用しようとした演技未経験者はフィードバック遅延により吃音症状が引き起こされたことで自然性が低下し，演技経験者は自らのイメージに没入することでより自然な発話ができた」(仮説 2') ことを示唆する．新たなこの仮説が明確に示されれば，フィードバック遅延を改善することで，変換音声のフィードバックがキャラクタ性を付与するようなユーザの発話変容に有用と示されることが期待できる．特に提案システムに代表される音声フィードバックは演技経験のないユーザに有用となることが考えられ，結果 1: 表 2 で Usability の評価は演技未経験者が高い傾向にあったこととも矛盾しない．また，結果 1: 表 3 で no FB と chara FB の一人称 Similarity 評価において見られた chara FB 有意な傾向が，より明確なものとなることが期待される．

5. まとめ

本稿では，変換音声に所望のキャラクタ性を付与するようにユーザを発話変容させることを目的として，リアルタイム DNN 音声変換によるフィードバックシステムを提案した．実験的評価により，演技経験の少ないユーザに対して特に有用性が高いことや，F_0 のみを変換することで十分な発話変容効果があることなどが示唆された．同時に，フィードバック遅延による発話の自然性低下が第一の課題として確認された．遅延時間を改善した後，日常的に目標話者の発話や提案システムに慣れ親しんだユーザのシステムへの適応過程を調査し，本研究によって得られた新たな仮説を検証していくことを今後の展望としたい．

謝辞：本研究開発は総務省 SCOPE(受付番号 182103104) の委託を受けて実施した．

参考文献

[1] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[2] 峯松信明, “音声分析・合成・認識技術を用いた多様な外国語教育支援,” **日本音響学会誌**, vol. 74, no. 9, pp. 525–530, Sep. 2018.

[3] M. Lincoln, A. Packman, and M. Onslow, “Altered auditory feedback and the treatment of stuttering: A review,” *Journal of Fluency Disorders*, vol. 31, no. 2, pp. 71–89, 2006.

[4] T. Toda, A. W. Black, and K. Tokuda, “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.

[5] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3893–3896.

[6] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4869–4873.

[7] R. Arakawa, S. Takamichi, and H. Saruwatari, “Implementation of dnn-based real-time voice conversion and its improvements by audio augmentation and mask-shaped device,” in *Proc. SSW*, Vienna, Austria, Sep. 2019, pp. 93–98.

[8] T. Saeki, Y. Saito, S. Takamichi, and H. Saruwatari, “Real-time, full-band, online DNN-based voice conversion system using a single CPU,” in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 1021–1022.

[9] D. Ladd, K. E. Silverman, F. Tolkmitt, G. Bergmann, and K. R. Scherer, “Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect,” *The Journal of the Acoustical Society of America*, vol. 78, no. 2, pp. 435–444, 1985.

[10] M. Fukuoka, “Japanese language learners’ recognition order of emphasis and paralanguage speech acts,” *Journal of the Phonetic Society of Japan*, vol. 78, no. 3, pp. 1–14, 2017.

[11] H. Ming, D. Huang, M. Dong, H. Li, L. Xie, and S. Zhang, “Fundamental frequency modeling using wavelets for emotional voice conversion,” in *Proc. IEEE ACII*, Atlanta, U.S.A., Sep. 2015, pp. 804–809.

[12] W.-C. Huang, T. Hayashi, Y.-C. Wu1, H. Kameoka, and T. Toda, “Voice Transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining,” in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 4676–4680.

[13] L. Zhang, C. Yu, H. Lu, C. Weng, C. Zhang, Y. Wu, X. Xie, Z. Li, and D. Yu, “DurIAN-SC: Duration informed attention network based singing voice conversion system,” in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 1231–1235.

[14] P. Denes and E. Pinson, *The Speech Chain, 2nd Ed.* W.H Freeman and Co., 1933.

[15] Sony, “WH-1000xm4: wireless headphones with noise-canceling features,” https://www.sony.jp/headphone/products/WH-1000XM4/feature_1.html, (参照 2021-1-30).

[16] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, “ATR technical report,” no. TR-I-0166M, 1990.

[17] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, “Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network,” in *Proc. IWAENC*, Tokyo, Japan, Sep. 2018, pp. 286–290.