Vol.2020-MUS-127 No.18
Vol.2020-SLP-132 No.18
2020/6/6

『転声こえうらない』利用者の基本周波数分析

堀部 貴紀[2,1,a)]　石原 達馬[2]　白井 暁彦[2]　森勢 将雅[1,b)]

概要：本研究では，「転声こえうらない」利用者の音声を対象に基本周波数や信号対雑音比の分析し，統計的な性質について調査した．一般的なボイスチェンジャーでは，収録者となりたいスタイルの声質の差や収録環境によって，声質変換の物理的・印象的品質が下がることがある．本研究の目的は，これらの性質を把握し，様々な音声に対して高精度なボイスチェンジャーを実現するため，Web で無料利用できるボイスチェンジャーサービスの長期運用を通して，利用者の音声の物理的特徴について分析を行い，利用者の音声傾向を調査することである．本分析では，全体の統計的な性質分析だけでなく，自己申告の性別に基づいた基本周波数の分析も行った．基本周波数の統計的な性質を解析の結果，全体では男性と示唆される音声サンプルが多い一方，開発や SNS における普及展開のタームごとで区切ると，女性利用者の比率が増えていることがわかった．また自己申告の性別に基づいた入力と組み合わせることで，潜在的な利用者の分類が可能になり，利用者にあわせたボイスチェンジャーサービスの改善の指針を得た．さらに信号対雑音比を解析した結果，収録環境は利用者の環境に依存するために原因の特定は難しいが，部屋の環境や，録音機材としてのスマホの持ち方について，理想の方法を利用者に提案するための分類が提案できた．

1. はじめに

自分の声を特定の他者の声に変換するボイスチェンジャーは，これまでにも研究がなされており [1–3]，一般でも利用可能なサービスも提供されつつある [4]．これらの技術では，特定の話者間での変換ではうまくいくものの，さまざまな要因により満足できる品質には至らないこともある．例えば，収録者の声質となりたいタイプの声質の差が大きいとき，声質変換後の品質が下がることがある [5]．また声質劣化の要因のひとつとして，収録時に混入する背景雑音も挙げられる．雑音が入っていないクリーンな音声であれば，音声分析法の分析失敗に伴う劣化を避けることができるだろう．しかし，多くの利用者がクリーンな音声を収録する環境が整っている場所で音声を収録しているとは限らず，空調などの雑音の混入が音声分析の精度や，変換音声の声質に影響する．

「転声こえうらない」は，グリー株式会社 GREE VR Studio Lab が https://vr.gree.net/lab/vc にて公開している Web ブラウザ上で利用できるボイスチェンジャーを実現するサービスである．

本サービスは 研究目的で公開されていることが明記されており，ボイスチェンジャー体験とシェア機能を通して，利用者の「なりたい自分」についての調査，特に声の変換品質を改善するためのデータを収集している．利用者のデータを個人に紐づけないで収集する代わりに，無料で楽しく気軽に利用できるサービス設計となっている．ボコーダーベースで作られており，声の高さ（pitch）とフォルマント（formant）を 10 秒間だけ変換することで多様な声に変換できる．

図 1 「転声こえうらない」スクリーンショット

なりたい声のスタイルをキャラクター画像とともに英

1 明治大学
Meiji Uniersity
2 グリー株式会社
GREE Inc.
a) ev180600@meiji..ac.jp
b) mmorise@meiji.ac.jp

Vol.2020-MUS-127 No.18
Vol.2020-SLP-132 No.18
2020/6/6

図 2 「転声こえうらない」Twitter でのシェアの様子

語・日本語ともに 13 種類 (おねえさん，両声類，ソプラノ，小学生，ヤミ声，おにいさん，男子中学生，ダンディ，ムッシュ，カワボ，マダム，ゴリラ，ダミボ) 用意し，Twitter でシェアできる設計となっており，それぞれ利用者の自己申告による声のタイプ (男性，中性，女性) を設定することで あらかじめ用意されたプリセット (pitch, formant) を設定して，ブラウザの WebAudio 経由で保存したサンプリング音声をボコーダーに渡してサーバ上で変換する．

「転声こえうらない」では，なりたい声のスタイルを多種類設定しており，特定の話者の発話に変換するタイプのボイスチェンジャーとは性質が異なる．多種類の声質変換を実現することで，どのようなタイプの声質変換の場合に声質が劣化しやすいかを計測しやすくなる．本研究では，Web ブラウザだけで利用できる無料のボイスチェンジャーサービスを通じて，収録音声をサーバに保存し統計的解析ができるようにした．その中で、統計的に容易な基本周波数や，利用者の収録環境を把握するため信号対雑音比について分析する．

2. 関連研究

声質変換システムや音声コーパスを作成するにあたって，話者の物理的特徴や信号対雑音比 (以下 SNR) について統計的な調査を行った事例を以下で紹介する．

声質変換における環境音に関して，Monisankha Pal らが 2016 年に行った研究 [6] では，ノイズの入った音声の変換結果が，実験した全ての声質変換システムで音質の高さを客観評価する指標である Perceptual Evaluation of Speech Quality（PESQ）が下がったことが報告されている．

Vassil Panayotov らが行った LibriSpeech [7] では，新しい音声コーパスを作成するにあたって訓練データとしてあるいはテストデータとして話者の音声を利用するために，2484 人の話者が研究に参加しサブセットに基づく発話を収録した．LibriSpeech をもとに Text-to-Speech のために構築された LibriTTS [8] では，LibriSpeech と同様に訓練データとテストデータとして，2456 人の話者が研究に参加し発話を収録した．

以上のような事例は存在するが，数万件以上の公開実験事例は少ない．また大量の匿名サンプルで得られた音声の集合的特性や統計的 SNR に関する調査について，公開された方法での事例は少ない．スマートスピーカーのような音声認識を目的とした場合は不特定多数の利用者の音声サンプルが取得可能になるが，ボイスチェンジャー目的での利用とは異なるデータとなる可能性がある．

3. 理論

本研究では「転声こえうらない」の不特定多数の利用者を対象とした音声変換サービスの改善を目的としている．「転声こえうらない」は森勢らによる音声分析合成システム WORLD [9] をベースに，リアルタイムで動作する音声合成エンジン RealWorld を独自に開発し，将来的な新サービスへの利用を検討している (`https://www.slideshare.net/vrstudiolab/vrsionup6-slideshare-156082977`)．

音声変換サービス改善のため，声優や音声データベースなどの録音環境や定められたシナリオではなく，実際の利用者の自由な音声を対象にする必要がある．利用者の物理的特性の把握は聞き手の印象を含めた品質向上に貢献できる可能性があり，サンプルとして得られる話者数が統計的信号処理および物理的精度を高める要因になる．

3.1 平均基本周波数

以下，本報告での用語及び手法の定義を行う．まず音声ファイルの基本周波数 $F0$ とスペクトル包絡 S_p を抽出した．$F0$ を用いて，録音されたシーケンスにおける平均基本周波数 $\overline{F0}$ を求める．有声音区間の $F0$ の要素数を N とすると，$\overline{F0}$ は以下の式により与えられる．

$$\overline{F0} = \exp\left(\frac{\sum_{n=0}^{N-1}(\log F0_n)}{N}\right) \quad (1)$$

式 (1) では，0 より大きい $F0$ について人間の音の高さの知覚が対数的な尺度であるため，対数をとって平均値を求めた．以上をもとに得られた値を分析した音声ファイルにおける $\overline{F0}$ とした．

3.2 信号対雑音比

本解析では，WORLD におけるスペクトル包絡推定法である CheapTrick [10] により得られたスペクトル包絡 S_p を用いて，信号対雑音比 P_{SN} を求める．得られたスペクトル包絡は，時間と周波数の軸からなる 2 次元配列になっている．

各フレームにおけるスペクトル包絡の総和を求めて 1 次元配列 [sum_list] に直す．これはそのフレームにおけるパワーに相当する．また，[sum_list] を昇順でソートした 1 次元配列を [sorted_sum_list] と呼ぶ．次に，1 次元配列 [sorted_sum_list] の要素数のうち，上位・下位それぞれ 10 %に当たるインデックスを求める．これを h・l と呼ぶ．インデックスとして利用するために round 関数を使用して整数の形に整える．要素数を N とすると，それぞれのインデックスは以下で表される．

$$h = \mathrm{round}(0.9 \times N) \tag{2}$$

$$l = \mathrm{round}(0.1 \times N) \tag{3}$$

ソートされたパワーのうち h 番目を収録された信号のパワー P_{s}，l 番目を雑音のパワー P_{n} とすると，P_{SN} は対数を用いて式 (4) で求められる．

$$P_{\mathrm{SN}} = 10 \log \left(\frac{P_{\mathrm{s}}}{P_{\mathrm{n}}} \right) \tag{4}$$

「転声こえうらない」で収録された音声には，全ての区間に音声が含まれるわけではなく，ある程度の無音区間が存在すると考えられる．有声区間においても瞬間的なピークでパワーを計測することは妥当とはいいがたい．そのため，全区間のパワーから上位・下位のそれぞれ 10%のパワーを信号・雑音のパワーであると仮定して SNR を計算している．なお，P_{n} については CheapTrick でスペクトル包絡を求める際に 0 にならないように補正されている．そのため，P_{SN} の値が無限大に発散する問題は生じない．以上をもとに得られた値を分析した音声ファイルにおける P_{SN} とした．

4. 方法

「転声こえうらない」は Google Cloud Platform (以下 GCP) によって実装されており，利用者がブラウザの WebAudio 経由で録音した音声を PCM 形式で 16 kHz，16 bits，モノラルに圧縮されたファイルが GCP ストレージに格納されている．分析では，WAV ファイル形式でデコードした上で，Librosa.load 関数でサンプリングレート 16 kHz の波形として取り出して使用した．そして音声の基本周波数・信号対雑音比分析には，WORLD を python 用に拡張した pyworld ライブラリを使用した．pyworld ライブラリでは，DIO [11] と呼ばれるアルゴリズムが使われている．

本分析では，サービス開始の 2019 年 7 月 1 日から 2020 年 3 月 31 日までに「転声こえうらない」を利用した利用者の音声ファイルを対象とした．対象となる音声ファイル数は 40,467 件であり，録音環境は各利用者の環境に依存するが，主にスマートフォンの内蔵マイクや PC マイクが想定される．

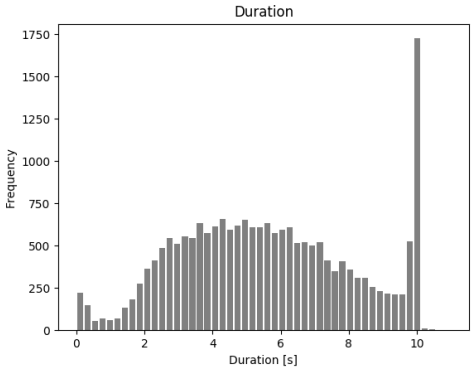

図 3 利用者の録音時間 [s] に対するヒストグラム

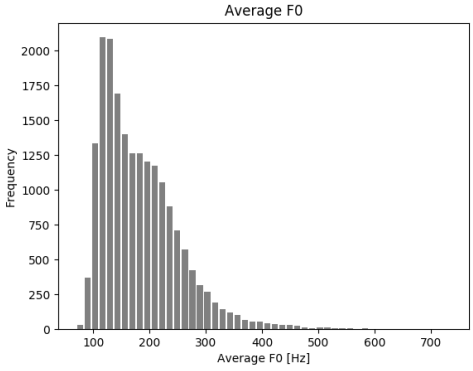

図 4 利用者の平均基本周波数 ($\overline{F0}$) [Hz] に対するヒストグラム

5. 分析結果

表 1 は，収録された音声全体の情報である．

表 1 分析結果

対象区間	2019 年 7 月 1 日から 2020 年 3 月 31 日
最大収録時間	10 秒
合計収録件数	40,467 件
重複を省いた件数	20,803 件

「転声こえうらない」では，同じ発話に対して異なるパラメータで変換できる．そのため，合計収録件数には，同じ発話が複数含まれることになる．その影響を省いた件数が 20,803 件である．

一度録音したサンプルの再利用を除外するため，本分析では，録音時間 Duration ・$\overline{F0}$・ P_{SN} が小数点 3 桁まで一致した音声ファイルを同一発話とみなして除外した．この条件に絞った 20,803 件のファイルを以後，全対象と呼ぶ．そのヒストグラムが図 3〜5 である．

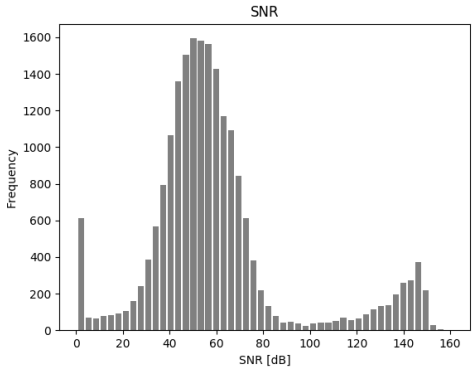

図 5 利用者の信号対雑音比 (P_{SN}) [dB] に対するヒストグラム

5.1 全対象の統計的性質

5.1.1 録音時間

全対象の録音時間をまとめたヒストグラムが図 3 である．まず制限時間上限である 10 s まで録音を行った利用者が多いことがわかる．図 3 では，372 件の 10 s を超えたファイルが確認できる．本分析でわかったことであるが，WebAudio の実装上ブラウザ上のタイマーと録音に多少のズレが発生することがあり，1.7% 程度発生することが確認できた．

また，録音時間が 0 s となった音声ファイルが全体の 1.9%程度存在した．これは利用者がマイクの取得を許可しなかったケースや録音開始直後にブラウザを閉じたケースであると考えられる．

5.1.2 平均基本周波数

平均基本周波数 $\overline{F0}$ についてまとめたヒストグラムが図 4 である．実験の結果，平均基本周波数が 100 Hz から 150 Hz の利用者が多いことがわかる．一般男性の基本周波数が平均 125 Hz と言われることから，本サービス開始されてからの約 8 ヶ月間において，男性利用者が多いことが推察できる．また，分析結果が 400~500 Hz となった音声ファイル（全体の 1.0%程度）について，視聴して検証したところ，利用者自身の発声ではなく，アニメ等の音源を流し込んでいるサンプルが多くあった．

5.1.3 信号対雑音比

信号対雑音比をまとめたヒストグラムが図 5 である．全体の 2.9%にあたる約 600 件が 0 dB であることが確認できる．今回の収録では波形全体で短時間パワーを求め，その上位と下位との比率で P_{SN} を算出しているため，0 dB であることは全区間のパワーが均一であることを意味する．今回の場合，収録時間全てにおいて無音であることや，何らかの原因により収録そのものが失敗していたことが原因と判断できる．

ヒストグラムでは，50 dB 付近にピークが観測される以外にも，150 dB 付近に，もう 1 つのピークが観測できる．この原因は，P_{SN} の算出法と収録における無音区間の扱いに起因する．一部の収録端末では，話声として観測されるよりも一定以上小さい振幅の波形に対し，振幅を完全に 0 とする機能が備わっている．スペクトル包絡を計算する際には，計算に用いる区間の振幅が完全に 0 の場合に備え，微小な雑音を加える処理を導入している．P_{SN} が 120 dB を上回る事例については，振幅が完全に 0 として計算された区間のパワーを雑音のパワーとみなすことで生じている．

50 dB を中心としたピークでは概ね正規分布に近い形状になっていることに対し，150 dB 付近のピークでは低い P_{SN} に向けて緩やかに減衰する傾向が観測できる．この原因は，話者の声そのものが小さいことや，マイクまでの距離が遠いこと，無音区間の振幅を 0 にする処理が散発的に実施されたことなどの複合要因によると考えられる．これらについては，個別の事例に対して詳細な検討が必要になる．

5.2 期間を区切ってまとめた分析

5.2.1 平均基本周波数

次に，3 ヶ月ごとにタームを区切って分析を行った．変換に用いられるプリセットの改善などサービス品質改善や，利用者とのコミュニケーション用公式 Twitter アカウント「@koeuranai」の担当キャラクターを 3 ヶ月ごとに変更している．この実験期間は四半期ごとに区切ってすすめていることから，2019 年 7-9 月 (T1)，2019 年 10-12 月 (T2)，2020 年 1-3 月 (T3) とする．

作成した $\overline{F0}$ のヒストグラムが図 6 である．件数ベースでは，各タームで減少推移しているが，$\overline{F0}$ の分布については，200 Hz 付近のピークの比率が増えていることから，T1 から T2，T3 にかけて女性比率が上がっていることがわかる．

5.2.2 信号対雑音比

3 ヶ月ごとに区間を区切って作成した P_{SN} のヒストグラムが図 7 である．環境面，失敗（0 dB），分布の傾向について P_{SN} では大きな変化は見られなかった．この結果は，F0 の変化では利用者の属性が変化していた一方で，利用者層の収録環境は概ね均一に分布していることを示唆する．

5.3 自己申告の性別を考慮した分析

本サービスでは，音声と性別を紐づけるため，2020 年 3 月から収録音声に対して自己申告で性別を男性，女性，中性の 3 種類の集計を始めた．そして，声のスタイルや自己申告の性別など詳細なデータを集計した 2020 年 3 月 1 日から 3 月 31 日までの 1 ヶ月間について性別に基づいた平均基本周波数のヒストグラムが図 8 である．

まず自己申告の性別で男性を選択した利用者の $\overline{F0}$ は，一般男性の平均 125 Hz 近辺が 1 番多いことがわかった．また，男性と申告した利用者の中にも，$\overline{F0}$ が 200 Hz~250 Hz 近辺という結果になった利用者もおり，一般女性の声の

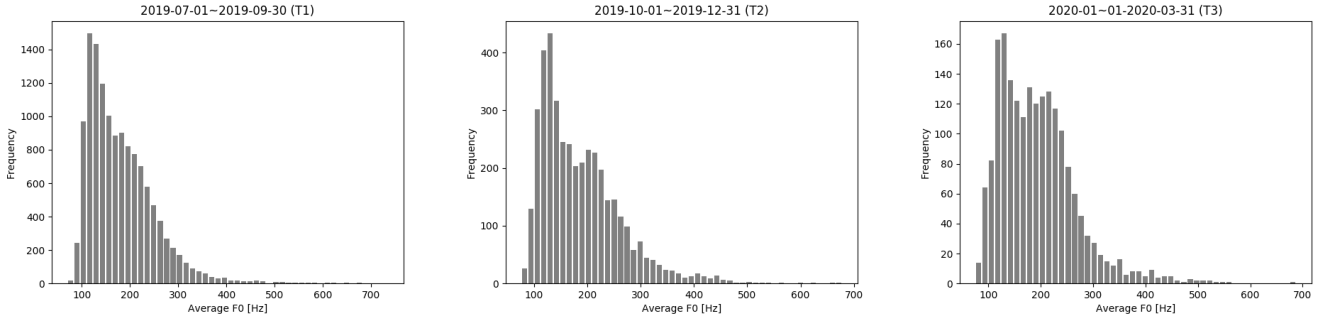

図 6 3 ヶ月ごとに分けた平均基本周波数 ($\overline{F0}$) に対するヒストグラム

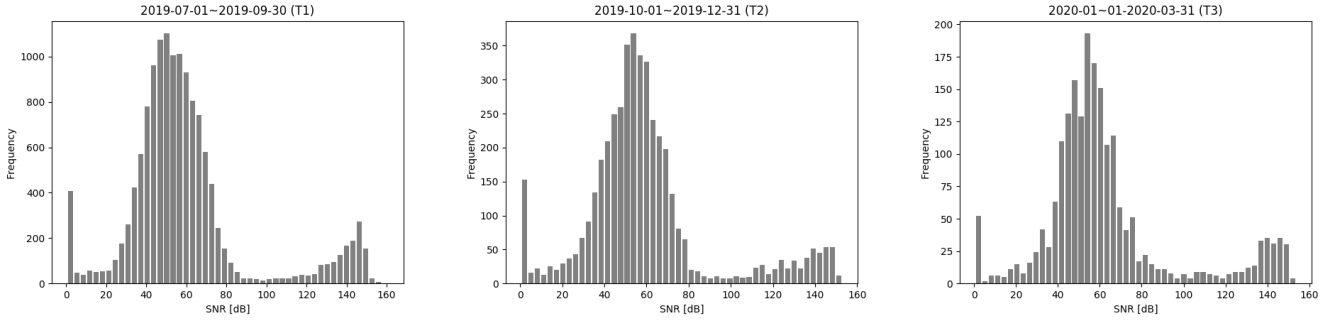

図 7 3 ヶ月ごとに分けた信号対雑音比 (P_{SN}) に対するヒストグラム

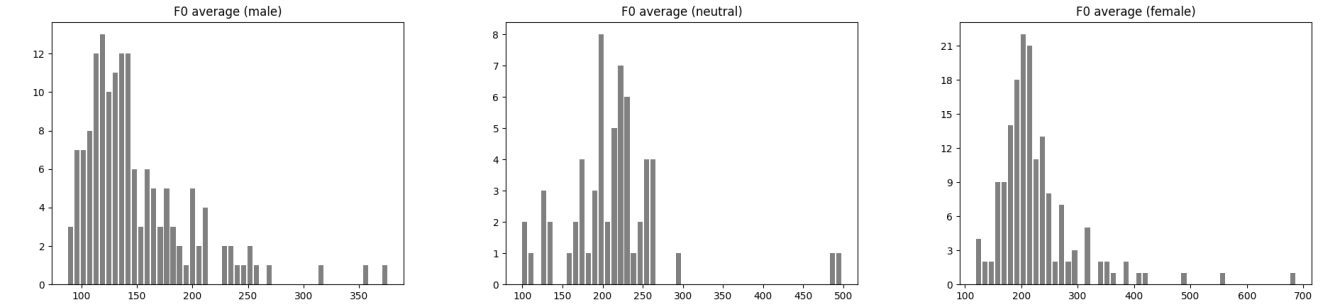

図 8 自己申告の性別ごとに分けた平均基本周波数 ($\overline{F0}$) に対するヒストグラム

高さのような特徴を持っていることがわかる．次に，自己申告の性別で中性を選択した利用者には，女性のように声が高い傾向にあることがわかる．しかし本分析のサンプル数が少ないため，今後も引き続き調査していく必要がある．

6. おわりに

本報告では，「転声こえうらない」利用者の平均基本周波数，信号対雑音比の分析をもとに，ボイスチェンジャー利用者の傾向を調査した．分析結果から，次のような事柄が明らかになった．

- 全体では男性と示唆される音声サンプルが多い
- 開発や SNS における普及展開のタームごとで区切ると，女性利用者の比率が増えている
- 自己申告の性別に基づいた入力と組み合わせることで，潜在的な利用者の分類が可能になり，利用者にあわせたボイスチェンジャーサービスの改善の指針を得られる

声質変換のパラメータを利用者の平均基本周波数をもとに決めることで声質変換の品質を向上させることができるだろう．今後の研究では，声質変換システムを利用者の平均基本周波数に合わせたパラメータを設定するアルゴリズムへ発展させられる可能性がある．

また，今回は選択した声のスタイルや自己申告の性別など詳細なデータの集計を続けることによって，どのような声の特徴をもつ利用者がどのような声のスタイルに変換したいのかを把握することが可能になるだろう．具体的にはユーザの物理的音声特徴から，好ましい変換（他のユーザに好まれる変換）へのリコメンドを行うことも可能であり，その前段となる分類器の構築に貢献できる可能性がある．

さらに，本分析による SNR 特性は，ボイスチェンジャーためだけでなく，Zoom 等の WebAudio 経由のハングアウト環境改善においても同様の特性をもつと考えられるため，遠隔相互授業や Webinar のためのソフトウェア設計や選定の指針としても利用できる可能性がある．具体的に

は，受講者の印象や多様な環境，端末環境を一つ一つ要素として調査するのではなく，WebAudio 経由で短いサンプル音声を提出させることで，無音区間と発声区間の利得を 50 dB 近辺として期待して標準化できる．これにより受講者の自然な発音や雑音環境を自動的に切り替える Push to Talk (PTT) 機能を実装した Webinar なども開発できるだろう．

さらに信号対雑音比を解析した結果，収録環境は利用者の環境に依存するために原因の特定は難しいが，利用者によりよい部屋の環境や，録音機材としてのスマホの持ち方について，指針が出せる程度の分類ができる可能性がある．

図 9 「録音スタイル」アンケートより

本実験の結果から現在では，サービスの「お問い合わせ」もしくは「音質の改善に協力する」というリンクから，画像とともに「録音スタイル」についてレポートを行うアンケートフォームを設置している．選択肢としては図 9 のように (内蔵マイクに口を近づける (低)，内蔵マイクに口を近づける (中)，内蔵マイクに口を近づける (高)，スマートフォンの画面を見る，電話スタイル，外部マイクを利用) といったスタイルを選択させることで，間接的なスマートフォン利用での理想の収録方法について知見を得ていきたい．データとしての分類や推定については明確な結果が得られていないため今後の課題とする．

また図 6 からは，利用者のペルソナについても推測ができる．Twitter ハッシュタグ「# こえうらない」では，公開されたユーザの変換結果をたどることができるが，女性の利用者であることが明確にわかる事例は少ない．一方で，全体のユーザ数の減少においても，女性比率が多くなっていることから，利用者のペルソナとして，「Twitter での公開は行わないが根強く使っている女性が多い」という見方もできる．

以上のように，数万件規模の無料ボイスチェンジャーサービスの音声サンプルを利用したサービス改善手法を報告したが，スマートスピーカーやボイスアシスタント等では，同様の手法を利用して，より多様な部屋環境の推定や，話者の特定や話者の位置推定，場所推定などが機械学習の連携で獲得できる可能性もあるだろう．

本研究のようにエンタテインメントもしくは自由な音声表現を対象とした研究は，スマートスピーカーのようなコマンドとは異なる情動にかかわる要素をとらえている可能性もあり，継続して調査を続けたい．

謝辞 本調査プロジェクト開発チームおよび，データ取得にご賛同いただいた皆様，特に調査の目的に賛同し「転声こえうらない」をご利用いただいた利用者すべてに感謝を記したい．

参考文献

[1] 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).
[2] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara: Voice conversion through vector quantization, ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 655-658 (1988).
[3] T. Toda, Y. Ohtani and K. Shikano: One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol. 4, pp. IV-1249-IV-1252 (2007).
[4] リアチェン voice, 入手先 ⟨https://crimsontech.jp/works/rcvoice/⟩ (閲覧日 2020/4/26).
[5] 河原英紀, 生駒太一, 森勢将雅, 高橋徹, 豊田健一, 片寄晴弘: モーフィングに基づく歌唱デザインインタフェースの提案と初期的検討, 情報処理学会論文誌, vol. 48, no. 12, pp. 3637-3648 (2007).
[6] M. Pal, D. Paul, M. Sahidullah, and G. Saha: Robustness of Voice Conversion Techniques Under Mismatched Conditions, arXiv preprint arXiv:1612.07523, (2016).
[7] V. Panayotov, G. Chen, D. Povey and S. Khudanpur: Librispeech: An ASR corpus based on public domain audio books, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5206-5210 (2015).
[8] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen and Yonghui Wu: LibriTTS: A corpus derived from LibriSpeech for text-to-speech, Proc. Interspeech 2019, pp.1526-1530, (2019).
[9] M. Morise, F. Yokomori, and K. Ozawa: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884 (July 2016).
[10] M. Morise: CheapTrick, a spectral envelope estimator for high-quality speech synthesis, Speech Communication, vol. 67, pp. 1-7 (March 2015).
[11] M. Morise, H. Kawahara and H. Katayose: Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech, AES 35th International Conference, CD-ROM, London UK (Feb. 11-13, 2009).