声に対する印象を用いた合成音声ライブラリ探索サービスの提案

清水　洸世　　梶　克彦
愛知工業大学

1　はじめに

人の歌声や喋り声を人工的に再現する音声合成ソフトは数多く存在しており，それらソフトのほとんどが複数種類の声を切り替えて使用できる．また，その中でもいくつかのソフトでは個人が声の元となる合成音声ライブラリを作成し，第三者による利用を前提とした配布を行える．例えば，喋り声を対象とした合成音声ソフト COEIROINK ではユーザの作成した音声合成モデルが 350 キャラクタ分以上配布されているほか [1]，歌声を対象とした合成音声ソフト UTAU では同ソフト上で使用できる UTAU 音源ライブラリが 7000 キャラクタ分以上存在する [2]．このように，今や合成音声ソフトの利用者は使える声に対し非常に多くの選択肢を持っており，その全ての把握は現実的ではない．

合成音声を利用するシーンにおいて，声が持つイメージや印象は声を選ぶ上で考慮すべき要素である．例えば喋らせるアナウンスの内容や，歌わせる曲調など用途に合った声質を持つライブラリを選択し，用いるのは制作において重要なプロセスである．しかし，現状声の持つ印象を知るには実際に聴いてみるのが最も有力な手段であり，数多あるライブラリの生み出す声を十分な数聴き比べ適切な声を採択するには多大な手間と時間を要する．その結果として，ユーザがライブラリを選ぶ際，その多くが普段の生活の中で聞いた経験のある声や，知っているキャラクターの声を選択していると考えられる．これはユーザ全体の中で使われる声に大きな偏りを生じさせてしまう．万に近い数存在するライブラリのうち実際にユーザに用いられる声は一握りであり，ほとんどのライブラリはユーザに用いられず埋もれてしまう．

図 1: サービスの画面イメージ

そこで本研究では，ライブラリごとの声に対する印象を事前に数値化し，それを用いてユーザの求める声に近いライブラリを探索するサービスを提案する．探索対象とする合成音声ソフトは特に利用できるライブラリが多く，後述する声質に関するアンケートが存在している UTAU 音源ライブラリを対象とする．本サービスでは事前に，声に対する印象を複数の印象軸ごとに評価スコアとして数値化する．ユーザは理想としてイメージする声の評価スコアを入力し，目的に合った音源を探索できる．評価スコアの軸には，例えばスコアの高低を女性らしい声・男性らしい声に対応させた”性別感”など，ユーザが声からスコアを，あるいはスコアから声をある程度想定できるような直感的な軸が望ましい．各音源に対する評価スコアは，アンケート調査によって集められたデータをもとに音源ファイルから各評価スコアを推定できる機械学習モデルを作成し，それを用いて付与する．また，本サービスは多くの人が手軽に利用できるよう Web サービスとして実装する．

2　関連研究

2.1　歌声に対して印象を推定する研究

本研究に関連する研究として，人間の歌声の情報を理解しやすい形で可視化する研究 [3][4] がある．この研究では，ある程度の長さがある歌声と，そのうちの瞬間的な長さの歌声を用い，それぞれで印象を推定するモデルを作成している．ある程度の長さがある歌声では，迫力性，丁寧さ，明るさの 3 軸に対してそれぞれのスコアを推定するモデルを作成し，推定を行なった．結果として，人の間でも印象の評価が大きく揺れるような歌唱などの例外を除き，十分な精度で印象を推定できた．また歌声のうち瞬間的な音声からは，印象に加えて声に対してイメージされる色を推定する試みも行っている．彩度など色の要素と表現語の一つとして挙げられた活動性との関係が見られるなど，声質の色での表現に対して一定の有効性が示されたものの，他の要素との関係については今後の課題とされている．この研究から，声質へのなんらかの評価軸を用いた評価スコア付与が可能である点や，それを用いた声質の表現は一つの手段として妥当なものであると考えられる．

2.2　合成音声の歌声にスコアを付与する研究

UTAU 音源を対象に声質に対し評価スコアを付与し，そのスコアを用いて音源を探索するシステムを提案する研究 [5] が存在する．この研究でも本研究と同じく UTAU に評価スコアを付与，探索システムを構築し，実際にユーザが求める声を探索できるかを確認している．スコアの推定には UTAU を用いて合成された音声データを用い，重回帰分析とカーネル回帰分析での推定制度の比較を行なっている．また，推定されたスコアを用いて音源を探索する際には，ある 2 つのスコア行列間のユークリッド距離を目標類似距離とし，その逆数として定義した目標類似度を用いて音源を探索している．評価実験では，ユーザのイメージする声に近いスコアを入力し，目標類似度の高いライブラリを提示した．するとユーザが求めるような声を持つライブラリを探索できる結果が示された．一方でこの研究では，探索システムの実装に留まっており，探索アルゴリズムの検討や実際にユーザが利用できるサー

ビスの提案は行われていない．

2.3 可視化・探索するシステムを提供するサービス

クリプトン・フューチャー・メディア株式会社と産業総合研究所によって開発された音楽発掘サービス「Kiite（キイテ）」では、その機能の一つとして「Kiite Radar」が提供されている [9]．ユーザは，楽曲の知名度やニコニコ動画上でのマイリスト率，楽曲を歌っているキャラクターなどの一般的な楽曲情報での絞り込みに限らず，楽曲の解析によって得られた曲の声質や踊りたくなるかどうかの印象をスライダーで設定し，その条件に合致する楽曲を探索するできる．加えて，全ての楽曲を分析した印象に基づいて 2 次元平面上にプロットされた印象マップが提供されている．激しい曲は左上に，軽快な曲は右上にプロットされるなど楽曲の配置には一定の傾向があり，ユーザはこれらを用いて直感的に楽曲を探索できる．対象が歌声でなく楽曲である点は異なるが，本来数値ではない印象を数値化し，それを用いて可視化・探索するという点で本研究と共通する部分があり，サービスを提供する上で参考にできる．

3 評価スコアを推定するモデルの構築

本研究では，UTAU 音源ライブラリに対して声質に対する評価スコアを付与するための機械学習モデルを作成する．それぞれのライブラリの音声から抽出した特徴量と，事前にアンケート調査などによって得られた評価スコアを用い，特徴量から評価スコアを推定するモデルを作成する．評価スコアの推論には学習データとして UTAU 音源声質アンケートのデータを，モデル作成には Python ライブラリである PyCaret を用いた．

3.1 特徴量の抽出

特徴量を抽出するためには，ライブラリを用いて複数の音階で「あ」「い」「う」「え」「お」「ん」と発声した音声ファイルを利用する．五母音と子音「ん」は連続して発声のできる音素であり，安定して音声の特徴を抽出できる．音階は A3，D4，G4，C5，F5 の 5 音階を用いる．複数の音階を用いる特徴量の数を増加させるほか，音声の音域による特徴への影響などを反映できると考えられる．

実際に用いる特徴量としては，MFCC，ZCR，F1〜F4 の周波数を採用した．MFCC は音声の周波数成分を人間の聴覚特性に基づいて変換したものであり，話者認識になどに広く用いられている．音声の声質に関連すると考えられるため採用し，本研究では MFCC を 64 次元まで取得し用いる．ZCR は音声の振動数を表す指標であり，音の高さやノイズの大きさによって振動数は変化する．音声の声高を固定した今回の音声であれば，音声の声質に関連する指標として扱えると考え採用した．音響フォルマントは音声の共振周波数を表す指標であり，母音の識別に用いられる．同じ音素のフォルマント周波数でも話者の年齢や性別によって変化があり [8]，声の印象に影響があると考えられる．本研究では各音階と音素ごとの F1〜F4 の周波数を取得し用いる．

3.2 モデルの構築

モデルの構築には Python ライブラリである PyCaret を用いた．PyCaret は機械学習モデルの作成を簡略化するためのライブラリであり，データの前処理からモデルの選定，評価までを一連の流れで行える．UTAU 音源声質アンケートの結果は Excel ファイルとして提供されている．このファイルに記載されている 240 種の UTAU 音源ライブラリのうち，現在でもダウンロードが可能であり，必要な音素が収録されていて，かつ利用規約上で研究目的を含む機械学習用途での利用が禁止されていない 168 ライブラリを対象とした．各評価スコアを推論の目標として与え，先述の特徴量から評価スコアを推定するモデルを評価スコアの 7 軸それぞれについて別々に作成した．学習時には，使用するライブラリから 3 割をランダムに選びテストデータとして，残りのデータを学習データとして用いた．学習モデルは PyCaret で選択できる手法のうち，連続モデルの中で学習後の R2 値が高かった AdaBoost Regressor を選択した．これは連続量である特徴量から連続量である評価スコアを推定するため，離散モデルよりも連続モデルの方が適していると考えたほか，離散モデルでの学習を試した際に過学習と見られる状況が多く見られたためである．

3.3 UTAU 音源ライブラリと UTAU 音源声質アンケート

本研究において探索対象とする UTAU 音源ライブラリは，無償で公開されている歌唱用音声合成ソフトである UTAU 上で使用できる音源ファイルであり，ソフトと同じく無償で公開されているものが多い．UTAU は波形接続と呼ばれる手法を用いており，音声データを切り貼りして音声を合成する．そのため，UTAU 音源ライブラリは主に合成に用いるためにできる限り一定の音程と音量になるように収録された収録者の肉声が収録されている．収録形式には複数の手法があり，単独音であれば各 wav ファイルにひらがな 1 文字に対応する音素が収録されており，連続音（VCV）であれば「あんああいあうあ」[10] といった形で複数の音素が連続して収録されている．

本研究で実際に特徴量を抽出する際，音声合成は行わずこの wav ファイルを直接操作し音声合成の手間を削減した．必要な音素が収録された音声ファイルのうち，UTAU においては子音部とブランクとして指定されるタイミング間の音声を利用した．この範囲は UTAU 上で音声が合成される際母音を伸ばすために用いる範囲であり，一般に安定して同じ声が収録されているため音声の特徴を抽出する上で適していると考えられる．複数音階の音声の生成には，音階ごとにライブラリで指定される wav ファイルからピッチシフトを行い，音声を生成した．

UTAU 音源声質アンケートはニコニコ大百科上で提言された UTAU 音源ライブラリに対する声の特徴を評価するためのアンケート規格であり，現在までにこの規格を用いて 250 種以上の UTAU 音源ライブラリに対してアンケートが行われている．このアンケートは表 1 に示す 7 項目について，それぞれ 1 から 7 までの 7 段階評価で 10 件以上のアンケート調査を行い，その平均を評価値としている．アンケートは各 UTAU 音源ライブラリごとに行われ，表情音源が複数存在する場合は各表情音源ごとに独立してアンケートが行われている．

Table 1: UTAU 音源声質アンケートの評価軸

評価軸	低い値の示す表現語	高い値の示す表現語
声の性別	女性的	男性的
滑舌	舌足らず	はきはき
特有性	素直	癖がある
声の年齢	幼い	大人びた
透明感	ノイジー	クリア
声の強さ	優しい	力強い
声の明度	暗い	明るい

UTAU 音源ライブラリは数が多い点だけでなく，声質の依存先

が生の音声ファイルである点や，音素情報を関連づける ini ファイルが一般的に用いられる形式であるため非常に扱いやすい点，UTAU 音源声質アンケートが存在する点が評価できる．無償で利用できるものがほとんどである点も，多くの合成音声の中から用いたい声を探したい利用者のニーズに合っている．ライブラリの声質を評価する上で，また利用者からの目線においても都合が良いと考え，本研究の対象として選定した．

3.4 結果の評価

構築したモデルによる推論の精度を評価するため，テストデータに対して推論を行い得た値と実際のアンケートによって得られた値を比較した．テストデータは先述の通り学習データからランダムに選ばれたデータであり，対象とした 168 ライブラリの 3 割にあたる 51 ライブラリを用いた．7 つの評価軸それぞれについて，テストデータでの精度を評価した結果を図 2，図 3 に示す．

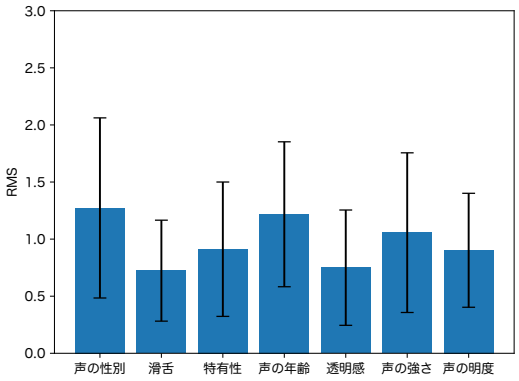

図 2: テストデータとの誤差の二乗平均平方根

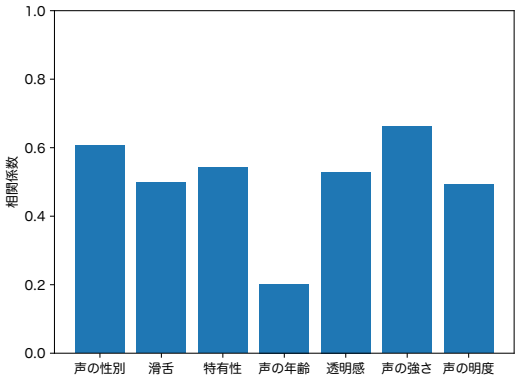

図 3: テストデータとの相関係数

図 2 では，横軸は評価軸を，縦軸はテストデータにおける実際の値と推測された値との誤差の二乗平均平方根（RMS），エラーバーは誤差の標準偏差を示している．RMS は実際の値と推測された値の誤差を示す指標であり，誤差が小さいほど推論の精度が高いと言える．結果として誤差の RMS は最も高い声の性別で 1.3 となった．これは先行研究 [6] の精度に近く，先行研究にて示された音響特徴量を用いたモデルの精度が再現できたと言える．

図 3 では，横軸は評価軸を，縦軸はテストデータにおける実際の値と推測された値との相関係数を示している．相関係数は一般に 0.7 以上であればデータ間に強い相関が，0.4 以上であればある程度の相関があるとされ，実際の値と推測された値における相関が強いほど推論の精度が高いと言える．結果を見ると，最も低い声の年齢は 0.20 とかなり低いものの，他の評価軸では 0.49 から 0.66 程度，最も高い声の強さでは 0.66 と，ある程度の相関が見られた．

相関係数の最も高い声の強さと最も低い声の年齢において，実際の値と推測された値の散布図を図 4 に示す．図を見ると，声の強さは右肩上がりの傾向が見られ，相関の存在が分かる一方で，声の年齢は点が縦に長く分布しているのが分かる．これは相関がないことに加え，実際の値はスコア範囲中に広く分布しているにも関わらず，予測された値がおおよそ 3〜5 と範囲の中央に偏っていることを表している．この傾向は程度の差はあれど声の性別と声の強さを除く全ての評価軸においても見られ，精度を下げている一因となっている．

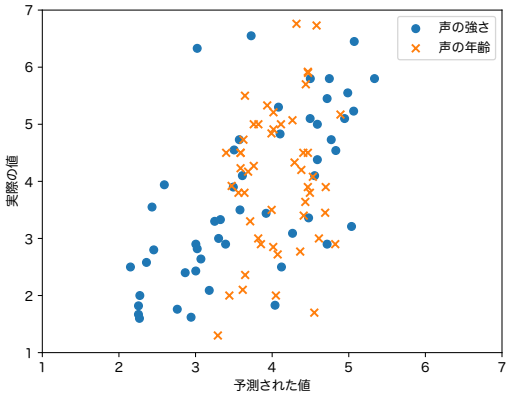

図 4: 実際の値と推測された値の散布図

予測値が中央に偏る現象について，一般にはデータの偏りやモデルの過学習，特徴量の不足などが要因とされている．学習データを確認したが，テストデータのスコア分布からも分かるように，実際のスコアは広く分布しておりスコア的な偏りは認められなかったほか，モデルの過学習についてもそのような傾向は見られなかった．一方で，今回用いた音声特徴量が声質の印象に対し十分でなかった可能性がある．主に母音の音素に対して音階ごとに別々の特徴量を抽出したが，音素間の遷移の特徴などより複雑な特徴量も声の印象に影響していると考えられる．そのほか先行研究 [6] では機械学習の手法としてディープラーニングのモデルが採用されていたように，学習手法は複数のものが考えられる．様々な手法を試し，より適したものに変えることで性能向上の余地があると考えられる．

4 声色見本帳

4.1 サービスの概要

本研究で提案するサービス「声色見本帳」は，ユーザが求める声の評価スコアを入力し，その評価スコアに近い合成音声ライブ

ラリを探索・提案するサービスである．このサービスを利用すると，ユーザは自身が求める声に近いライブラリを探すことができ，多くのライブラリの中から自身に合った声を見つけることができる．サービスは Web サービスとして実装し，ユーザはブラウザ上で利用できる．

4.1.1 ライブラリ探索

ユーザは求める声の評価スコアを 1〜7 の間で入力し，その評価スコアに近い UTAU 音源ライブラリを探索できる．ユーザは先に述べた 7 項目のうち、任意の数の評価軸に対してスライダーを操作し，評価スコアを入力する．必ずしも全ての評価軸に対して評価スコアを入力する必要はなく，探索の際重視しない評価軸についてはチェックボックスのチェックを外すことで評価軸を無効化し，その軸を無視して探索を行える．サービスは入力された評価スコアに近いライブラリを探索し，その結果をユーザに提示する．評価スコアが近いかの評価には，ユーザに選択された評価軸での平方ユークリッド距離によって判断している．登録された全てのライブラリに対して距離の計算を行い，その昇順にライブラリを表示し求める声に近いライブラリを上位に表示する．

4.1.2 ライブラリ詳細の表示

ユーザが探索結果として提示されたライブラリを選択すると，そのライブラリの詳細情報を表示する．詳細情報にはライブラリの名前，推定された評価スコア，サンプル音声，ライブラリの公式サイトや配布ページへの URL などが含まれる．サンプル音声としては，童謡の「かえるのうた」を Python ライブラリ「ScoreDraft」を用い事前に生成したファイルを用いる．本サービスはユーザが求める声に近いライブラリの探索のみを目的としているため，直接ライブラリをダウンロードする機能は提供せず，ユーザが自身で URL 先からダウンロードするよう促す．また，選択したライブラリの評価スコアを検索欄に転写できる．この機能を用いれば，あるライブラリに近い声質のライブラリを探したり，探索結果ページを送り反対に遠い声質のライブラリを探せる．

4.1.3 探索対象ライブラリの追加

ユーザは UTAU 音源ライブラリをアップロードし，サイト上に存在しない UTAU ライブラリを自由に追加できる．すでに構築されたモデルを用いて評価スコアやサンプル音声などを自動で生成し，データベースに追加する．ユーザがライブラリを追加できるため，より多くのライブラリや，この先新しく生まれ配布されるライブラリに対してもサービスに登録できる．

4.2 提案サービスの評価実験

(1) サービスを実際に使ってもらい，ユーザが求める声に近いライブラリを探索できるかを評価する．
(2) また，サービスの使いやすさや機能の拡張性についても評価する．

5 おわりに

本研究では，まずアンケートによって得られた評価スコアを目標とし，音声から特徴量を抽出してスコアを推定する機械学習モデルを構築した．その結果，声の強さなど一部の評価軸については相関係数 0.66 と一定の精度で推定できたものの，声の年齢など他の評価軸では予測値が中央に偏るなどの問題が見られた．構築したモデルを用いて，ユーザが声の印象をスコアとして入力し，それに近いライブラリを探索できる Web サービス「声色見本帳」を実装した．気軽に利用できる Web サービスを通じて目的に適した声との出会いを促進し，求める声質を持つライブラリを効率的に探索できたり，埋もれがちな多様なライブラリへ活用機会の増加が期待できる．

今後の課題としては，機械学習モデルの改善が挙げられる．3.4 節で述べたように音響特徴量と学習手法の選定による改善のほか，運用されるサービスの活用も考えられる．例えば，ユーザの探索履歴を収集し，あるユーザの探索パラメータを実際にダウンロードページにアクセスされた UTAU 音源の評価に影響させたり，直接ユーザに評価の投票を促しその結果を用いて追加で学習を行うなど，ユーザからのフィードバックを利用した推定精度の向上手法も考えられる．サービスとしての利便性向上も課題である．探索時に用いるライブラリ間の評価スコアの類似度指標の改善や，ライブラリ情報の更新・修正機能の実装，また権利者からの削除依頼への適切な対応体制の整備などが考えられ，これらの改善を行うことでより利便性の高いサービスを目指す．

参考文献

[1] COEIROINC，MYCOEIROINK．https://coeiroink.com/mycoeiroink/list

[2] Vocaloid Database．https://vocadb.net/Search?searchType=Artist&artistType=UTAU

[3] 金礪 愛，中野 倫靖，後藤 真孝，菊池 英明，歌声の印象評価尺度の構築に基づく多様な印象の自動推定手法．情報処理学会論文誌，Vol.57，No.5，pp.1375-1388，2016

[4] 金礪愛，アマチュア歌唱者に向けた歌声可視化方法の検討．博士論文，早稲田大学，2018

[5] 山根壮一ほか，歌声合成システムの音源データに対する声質推定と声質制御．情報処理学会研究報告，Vol.2015-MUS-108，No.6，pp.1-6，2015

[6] 横森文哉，大柴まりや，森勢将雅，小澤賢司，スペクトル包絡情報を入力とした Deep Neural Network に基づく歌声のための声質評価情報処理学会音楽情報科学研究会，Vol.2015-MUS-107，No.61，1-6，2015

[7] 佐賀 圭真，井村 誠孝，発話音声の聞き取りやすさ向上のための音声特徴量解析．エンタテインメントコンピューティングシンポジウム 2019 論文集，2019，pp.84-86，2019

[8] 粕谷英樹ほか，年令, 性別による日本語 5 母音のピッチ周波数とホルマント周波数の変化，日本音響学会誌，24 巻，6 号，pp.355-364，1968

[9] 産業総合研究所，音楽印象分析・音楽推薦を駆使して楽曲と出会える音楽発掘サービス「Kiite」を公開 https://www.aist.go.jp/aist_j/press_release/pr2019/pr20190830/pr20190830.html

[10] 巽式 連続音の録音リスト配布 - 巽のブログ，https://tatsu3.hateblo.jp/entry/ar426004

[11] ニコニコ大百科，UTAU 音源声質アンケートとは，https://dic.nicovideo.jp/a/utau%E9%9F%B3%E6%BA%90%E5%A3%B0%E8%B3%AA%E3%82%A2%E3%83%B3%E3%82%B1%E3%83%BC%E3%83%88