声に対する印象を用いた合成音声ライブラリ探索システムの提案

情報　太郎　　情報　花子
情報大学情報学部

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り，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の選定は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こと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．さらにその結果として，多くのユーザがライブラリを選ぶ際，普段の生活の中で聞いた経験のある声の中から声を選択し，結果としてユーザ全体の中で使われる声に大きな偏りが生じる問題も発生する．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なく埋もれてしまう．

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システムを提案する．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．本システムでは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し，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することで，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，本システムは多くの人が手軽に利用できるよう Web アプリケーションとして実装する．

2　関連研究

2.1　アマチュア歌唱者に向けた歌声可視化方法の検討

本研究に関連する研究として，人間の歌声から印象やイメージされる色を推定する研究 [3] 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，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．

2.2　声質を可視化する研究

本研究と同じく 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 [4] 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結果，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ことが示された．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することを想定したシステムの提案は行われていない．

3　*声質に対する評価スコアの推定

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．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　UTAU 音源ライブラリと UTAU 音源声質アンケート

まず，今回用いる UTAU 音源ライブラリについて説明する．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することで音声を合成する．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」[5] といった形で複数の音素が連続して収録されている．音声に対応する文字列（エイリアス）と音声の対応は oto.ini ファイルによって定義されており，UTAU はこのファイルを参照して子音や母音の開始位置を把握し合成に利用する．一部のライブラリでは，掠れ声や甘い声など，声質を意図的に変化させたものや，

違う音程で追加収録したものが存在し，それぞれ表情音源，多音階音源と呼ばれる．

次に，UTAU 音源声質アンケート [6] について説明する．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，声の性別，滑舌，特有性，声の年齢，透明感，声の強さ，声の明度の 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．

このように，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，また音素情報を関連づける ini ファイルも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などから，ライブラリの声質を評価する上で都合が良いため，本研究の対象として選定した．

3.2 音響特徴量の抽出

UTAU 音源声質アンケートのデータを用いて，UTAU 音源ライブラリに対する評価スコアを推定する機械学習モデルを作成する．UTAU 音源ライブラリはその声のみをデータとして持つため，歌唱時のピッチ遷移や発音の癖などが存在せず，歌声に対する印象はその声質のみに依存すると考えられる．そのため本研究では UTAU 音源ライブラリ中の音声ファイルを直接操作して特徴量を抽出し用いる．特徴量の抽出には「あいうえお」の母音と「ん」の 6 つを対象とし，単独音音源の場合はそれぞれのエイリアスを，連続音音源の場合は「- あ」など無音から繋がるエイリアスを対象とする．取得したエイリアスに対応する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抽出する．

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．また，前処理としてそれぞれの音声を A3，D4，G4，C5，F5 の 5 音階にピッチシフトさせ 5 パターンの音声を作成する．ピッチシフト時の元の声高の推定は UTAU 音源ライブラリに同梱されることの多い frq ファイルを参照するものとし，それを基準に音声ファイルを変換する．frq ファイルが存在しない場合は，事前にフリーソフトである OpenUTAU を用いて frq ファイルを作成する．また，多音階音源を対象とする場合は，各声高ごとにライブラリ中で指定される音階のファイルを利用する．

特徴量には各音階の音声ごとに MFCC，ZCR，F1〜F4 の周波数を取得し用いる．MFCC とは，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，音声の特徴を抽出するために広く用いられる．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，音声の特徴を抽出するために広く用いられる．MFCC のうち高次元の成分は話者認識に用いられるなど，音声の声質に関連するとされる [] ため，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，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．母音の識別の易難は，声を聞いた時の印象に関わると考え，本研究では F1〜F4 の周波数を取得し用いる．

3.3 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．

評価スコアの 7 軸それぞれについて別々にモデルを作成し，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 で選択できる多数の手法の中から，それぞれの評価スコアに対して各手法で初期モデルを作成し，最も推論に適していると思われる手法を主に R2 値から判断して選定し，学習を行った．

3.4 *結果の評価

(1) 学習時のテストデータでの精度の評価を記載，またその精度がどの程度のものであるかについて考察する

7 つの評価軸それぞれについて，学習時のテストデータでの精度を評価した結果を表**??**に示す．

4 *ここに探索システムの名前を入力

4.1 *システムの概要

(1) システムの画面イメージを記載
(2) 本システムはユーザが欲しい声の評価スコアを想定し，入力することでその評価スコアに近い UTAU 音源ライブラリを探索するシステムである．
(3) スコアの直入力のほか，既存のライブラリからスコアを転写する類似検査機能を持つ．
(4) またユーザはシステム上に存在しない UTAU 音源を追加することもできる．追加時はローカルで音声ファイルを解析，評価スコアの付与を行い，システムに追加する．

4.2 *システムの評価

(1) システム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，システム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

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，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

[4] 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

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

[6] 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