

愛知工業大学情報科学部情報科学科
コンピュータシステム専攻

令和6年度 卒業論文

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に関する研究

2025年2月

研究者 K21066 清水洸世

指導教員 梶克彦 教授

目次

第 1 章	はじめに	1
1.1	背景	1
1.2	研究目的とアプローチ	1
1.3	本論文の構成	3
第 2 章	関連研究	4
2.1	歌声に対して印象を推定する研究	4
2.2	合成音声の歌声にスコアを付与する研究	4
2.3	音声を可視化し探索する研究・サービス	4
第 3 章	評価スコアを推定するモデルの構築	6
3.1	UTAU 音源ライブラリと UTAU 音源声質アンケート	6
3.2	特徴量の抽出	7
3.3	モデルの構築	9
3.4	結果の評価	9
第 4 章	声色見本帳	13
4.1	要求仕様	14
4.2	実装	14
4.2.1	ライブラリ探索機能	14
4.2.2	ライブラリ情報の表示	15
4.2.3	探索対象ライブラリの追加	15
第 5 章	おわりに	17
5.1	まとめ	17
5.2	今後の課題	17
	謝辞	19
	参考文献	20

第1章 はじめに

1.1 背景

近年、人の歌声や喋り声をコンピュータ上で再現する音声合成ソフトウェアは楽曲制作やアナウンスなどに広く利用されており、その普及と同時に音声合成ソフトウェアの数も増加している。合成音声ソフトウェアの多くは合成音声ライブラリを切り替えて合成される声の種類を変更でき、ユーザが利用できる声の種類はソフトウェアの数以上に存在する。加えて、いくつかのソフトでは個人のユーザが自分の声からライブラリを作成して第三者に配布でき、そのようなソフトではライブラリ数は非常に多くなる。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 [1] ほか、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、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いる。

合成音声を利用するシーンにおいて、声の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、重要な情報をアナウンスする際に滑舌の悪い声を使うと情報が正しく伝達できない可能性があり、可愛らしく軽快な楽曲に対して力強い声を使用すると聞き手に違和感を与える恐れがある。用途に合った声質を持つライブラリの採択は、情報を伝える効果や楽曲の表現力の向上に寄与する。しかし、現状声の持つ印象を知るためには試聴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に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、ユーザ全体の中で使われる声には大きな偏りが生じている。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埋もれてしまう。

1.2 研究目的とアプローチ

本研究の目的は、多くの合成音声ライブラリの中から、ユーザの用途に合った声・ユーザのイメージする声に近い声を探索する効率的な手法の、実際にユーザが使用可能な形式での提案である。このような手法が実現できれば、ユーザは声質に対する自身の要求を入力するだけで、膨大な数の合成音声ライブラリの中から要求に合った声を持つライブラリを効率的に採択できるようになる。これにより、ユーザは従来のように実際に多くの声を聴き比べる必要がなくなり、声の選択にかかる労力を大幅に削減できる。また、純粋にライブラリの持つ声質のみでの探索はこれまで埋もれていた多様な声質を持つライブラリの発掘にもつながり、合成音声ライブラリの制作者とユーザの双方への利益が期待できる。

本研究では、ライブラリごとの声に対する印象を数値化する機械学習モデルを構築し、それを用いてユーザの求める声に近い UTAU 音源ライブラリを探索する Web サービス「声色見本帳」を提案する。サービスの概要を図 1.1 に示す。サービスを利用するユーザはまず、用途に合った声や自分の求める声をイメージし、サービスで用いる 7 つの評価軸に基づいてその声に対する評価スコアを数値化する。評価スコアは評価軸ごとに声に対する印象を 1~7 の数値で表すもので、例えば「性別感」という軸では、スコアの高低を男性らしい声・女性らしい声に対応させるなど、スコアと声の関係を理解しやすい評価軸を用いている。例えば女性らしい声を求める場合、「性別感」軸のスコアを 2 など低い数値に設定する。求める声のイメージを評価スコアとして数値化し入力すると、サービスは登録されているライブラリのスコアと

入力されたスコアを比較し、声の近いライブラリを複数提案する。ユーザは求める声に似ていると提案されたいくつかのライブラリを聴き比べ、自分の求める声に近いライブラリを少ない手間で探索できる。

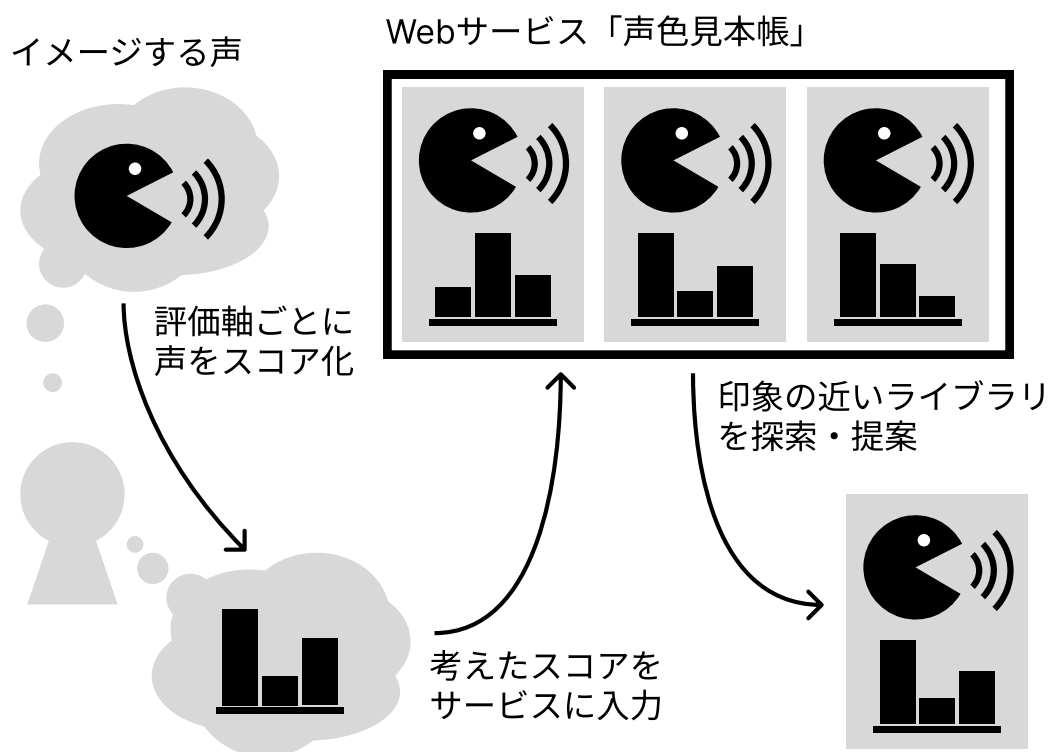


図 1.1: 提案するサービス「声色見本帳」の概要

提案するサービスでは、探索対象として登録されたライブラリごとの評価スコアを事前に設定する必要がある。声に対する印象を数値化するための評価スコアを得るには、アンケート調査の実施が方法として考えられる。しかし、アンケート調査には多くの時間と労力が必要であり、多くのライブラリに対してのアンケート調査は困難である。そこで、本研究では事前にいくつかのライブラリに対してアンケート調査を行ない、そのデータを用いてライブラリごとの声に対する評価スコアを推定する機械学習モデルを構築する。機械学習モデルを用いた評価スコアの推定の概要を図 1.2 に示す。構築したモデルを用いれば、アンケート調査を行っていないライブラリに対しても評価スコアを機械的に、一定の精度を持って付与できる。機械学習モデルを用い、より多くのライブラリを探索対象として登録できるほか、今後新たに公開されるライブラリに対しても効率的に評価スコアを付与できる。

本サービスでは UTAU 音源ライブラリを探索対象として選定した。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ウェアである UTAU、およびその互換ソフトウェア上で使用できるライブラリであり、このライブラリを切り替えると実際に合成される音声を変更できる。UTAU 音源ライブラリは 1.1 節でも触れたように数が特に多く、そのほとんどを無料で利用できるため、多様な声質を持つライブラリを探索するメリットが大きいと考えられるためである。また、UTAU 音源ライブラリには声質に関する調査を行なったアンケートデータが既存するため、機械学習モデルの構築に用いるデータセットの作成が容易である点も選定の理由である。なお、UTAU 音源ライブラリを利用し構築したモデルでも、他の合成音声ソフトの音源ライブラリに対して適用が可能であり、将来的に UTAU 音源ライブラリ以外のライブラリに対しても探索できるようサービスを拡張できる。

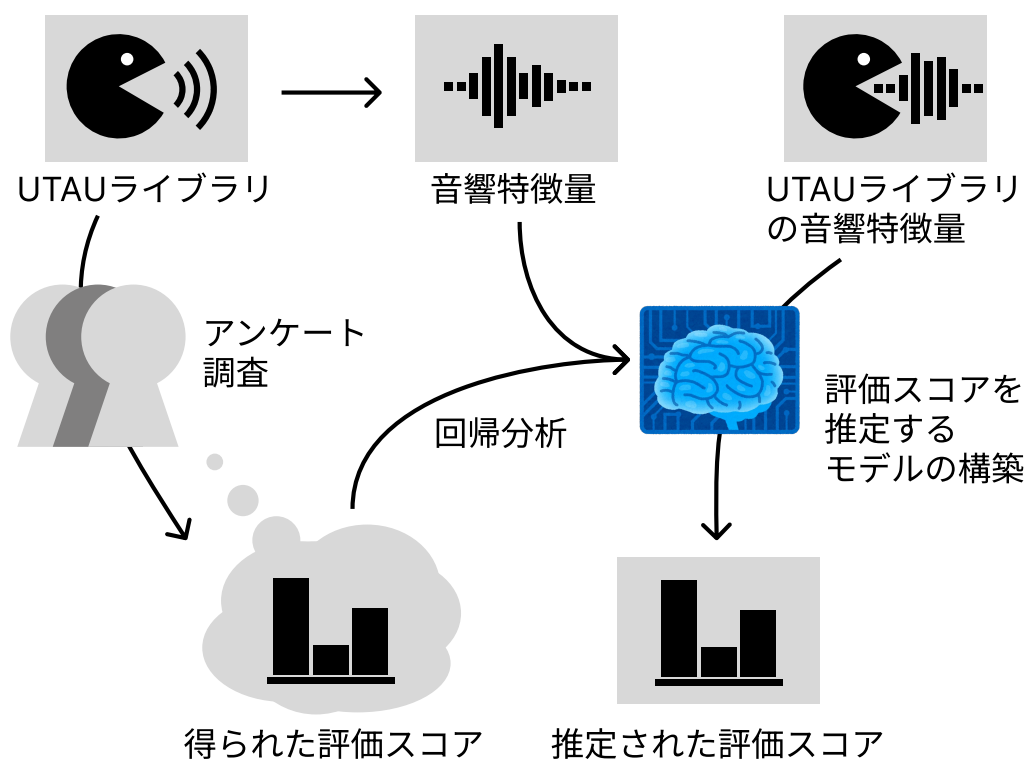


図 1.2: 評価スコアを推定するモデルの構築イメージ

1.3 本論文の構成

本論文の構成は以下の通りである。第2章では、関連研究について述べる。第3章では、ライブラリから評価スコアを推定する機械学習モデルの構築と、構築したモデルの推論精度に対する評価実験について述べる。第4章では、提案するサービス「声色見本帳」の設計と実装について述べる。最後に第5章として本論文のまとめと今後の課題について述べる。

第2章 関連研究

2.1 歌声に対して印象を推定する研究

人の歌声を対象に、その印象を推定する研究は数多く存在する。金礪らはアマチュア女性歌唱者の歌声を対象に、同じメロディを歌った歌声から音響特徴量を抽出し、「迫力性」「丁寧さ」「明るさ」の3つの評価語に対するスコアを推定するモデルを作成している [3]。この研究では、歌声から 1 msec ごとに音響特徴量を抽出し、評価語との関係を重回帰分析によってモデル化し、そのモデルを用いて評価語のスコアを推定している。また田中らの研究 [4] では、既存の楽曲から抽出した歌声を用い歌声の印象評価語を推定するモデルを作成している。この研究では歌声に対する印象評価語を楽曲ごとに付与した後、抽出した歌声を歌詞のフレーズごとに分割し音響特徴量を抽出、ランダムフォレスト法で学習した分類器を作成している。結果として、「透明感のある」という印象評価語に対して高い精度での推定を実現している。これらの研究から、歌声に対して印象を推定するための評価軸を用いた評価スコア付与が可能である点や、それを用いた歌声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。

これらの研究は実際に人が歌唱している音声から音響特徴量を抽出しているが、実際の歌唱には歌唱者の声質の差異以外にも、例えばビブラートの強さやしゃくりの癖、声区の使い方など歌い方の特徴が表れる。ビブラートやしゃくりはピッチの変動によって生じるテクニックであり、F0 の変動量に大きな影響を与える。他にも音響特徴量から声区の使い方を推定する研究 [5] があるように、歌い方は音響特徴量に影響を与えており、実際の歌唱から抽出した音響特徴量には歌い方の特徴が含まれていると言える。歌い方の特徴は歌声の個人性知覚に対して大きく影響を与える [6][7] ため、音響特徴量に含まれる歌い方の特徴は、歌声の印象推定にも影響を与えていると予想できる。一方で、本研究で対象とする UTAU を含め多くの合成音声において歌い方はライブラリではなくソフトの使用者に依存するため、それらの情報にまつわるライブラリごとの音響特徴量取得が不可能である。したがって、本研究では音響特徴量に含まれているビブラートやしゃくりなどの歌い方の特徴量に頼らない分析が必要となる。

2.2 合成音声の歌声にスコアを付与する研究

本研究でも扱う UTAU 音源ライブラリと UTAU 音源声質アンケート、あるいはその規格を用いた研究が存在する。山根らによる研究 [8] では、40 個の UTAU 音源ライブラリを用いて合成した音声データから声質特徴量ベクトルを抽出し、ベクトルからスコアを推定するモデルを重回帰分析とカーネル回帰分析を用いて作成している。スコアの推定に用いる音声データとして短音節発声、話声、歌声の3種類を用いた際の推定精度を比較しており、その結果として音声データの種類による推定精度の差異は少ないとされている。横森らによる研究 [9] では 80 個の UTAU 音源ライブラリから特徴量として MFCC を抽出し、DNN(Deep Neural Network) を用いて声質の評価スコアを推定し、評価実験を行なっている。これらの研究ではスコアを推定するモデルの作成に止まっており、そのスコアを用いたライブラリの探索システムやサービスの作成は行われていない。

2.3 音声を可視化し探索する研究・サービス

音声のある形式で可視化し、それを探索に用いる研究やサービスが存在する。佐治らは、喋り声に対して話者特徴量を抽出して可視化し、その視覚情報をもとに音声を取り出す検索システムを作成している

[10]. 32次元の値として得た話者特徴量を主成分分析によって2次元に圧縮し、2次元空間上にプロットし話者特徴量の分布を可視化している。話者特徴量は人間には直感的に理解しにくく、2次元空間上にプロットしてもその配置に一貫した意味があるかは不明である。しかしある程度似た声質を持つ話者は近い位置にプロットされる傾向があり、結果として話者の性別ごとにある程度固まってプロットされるため、これを用いた話者の探索が可能と分かった。また、同研究では2次元空間上へのプロットと声真似による入力を組み合わせた検索手法を提案している。これはユーザに地声と求める声の声真似を入力させ、それらの特徴量の差を「ユーザがイメージする声質へ向かうベクトル」と仮定し2次元空間上の範囲を限定する手法である。評価実験によってこの手法は探索時間の短縮が可能だと示されている。この研究は人にとって直感的に理解できない特徴量を用いているものの、その特徴量を2次元平面として可視化し、さらにユーザの声を入力に用いて探索を効率化している。

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 [11]。このサービスでは、楽曲を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できる。本研究で扱っている声質を楽曲の絞り込みに用いられるが、「かっこいい」と「かわいい」を両端に持つスライダー1つのみが用意されており、ユーザはこのスライダーを動かして楽曲を探索できる。加えて、全ての楽曲を分析した印象に基づいて2次元空間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。

第3章 評価スコアを推定するモデルの構築

本章では、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の概要と実装について述べる。3.1 節では推論に用いる UTAU 音源ライブラリと UTAU 音源声質アンケートについて述べる。3.2 節では使用する音声ファイルと音響特徴量、その抽出方法について述べる。3.3 節では評価スコアを推論する機械学習モデルの構築について述べ、3.4 節で構築したモデルの精度を評価する。

3.1 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 音源ライブラリと、声質評価の基準及び学習データとして用いる UTAU 音源声質アンケートについて説明する。UTAU 音源ライブラリは wav ファイルとメタデータを含むファイル群からなるデータセットであり、UTAU 音源ファイルとして提供される。wav ファイルの中身は収録方式によって多少異なるが、現在広く用いられている連続音方式であれば「あんあいうえあ」[15]といった形で複数の音素をできる限り一定の音程と音量になるように収録された肉声が wav ファイルとして保存されている。図 3.1 で実際の UTAU 音源ライブラリの一部を示す。個人が作成・公開が可能で、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、この音声データを切り貼りして加工し、歌声を合成する。

名前	変更日	種類
 _あいうえあ A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _いいうあえいえ A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _いいゆやいえいえ A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _いえいえゆよいえよ A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _ヴあヴあヴいヴあヴヴえヴあ A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _ヴいヴいヴヴあヴえヴいヴえ A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _ヴいヴいヴゅヴゃヴいヴいヴい A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _ういういうわういう A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _ヴいえヴいえヴゅヴよヴいヴよ A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _うういおあおい A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _ヴヴいヴおヴあヴお A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _うういうおわうおう A4.wav	2024年1月29日 19:05	WAVE オーディオ
 _うううううううう A4.wav	2024年1月29日 19:05	WAVE オーディオ

図 3.1: UTAU 音源ファイルの一部

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してその声の特徴を評価するためのアンケート規格であり [16]、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。このアンケート規格では、UTAU 音源の声に対して表 3.1 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するとされている。実際のアンケートの実施方法や具体的な集計件数は定められておらず、アンケートの実施者が各自で決め行っている。

表 3.1: 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

3.2 特徴量の抽出

UTAU 音源ライブラリから推論に用いる音響特徴量を抽出する方法について述べる．特徴量の抽出には，UTAU 音源ライブラリから複数の音階で「あ」「い」「う」「え」「お」「ん」と発声した音声ファイルを直接操作して作成し，利用する．これらの音素を選択した理由は，五母音と子音「ん」が継続的に発声可能な音素であり，安定した音声区間から音響特徴を効果的に抽出できるためである．音声ファイルの操作には，ライブラリに「固定範囲外」として指定される範囲を利用した．図 3.2 に「あ」を発声している音声ファイルの波形とスペクトログラム，ライブラリに指定されるタイミングの例を示す．この図は UTAU 互換ソフトウェアである OpenUTAU からキャプチャしたものである．波形上の白色の背景の範囲が固定範囲外であり，UTAU が実際に音声の合成を行う際は母音を伸ばすためにこの範囲の音声を加工している．固定範囲外は音素の発声始めと終わりの部分を含まないため，発声とひいては音声波形・スペクトログラムの形状が安定しており，音声の特徴を抽出しやすいと考え，この音声区間を用いた．音声の評価には各音素の発声を 5 つの音階ごとに生成した．歌唱において現実的に使用されたと考えた表 3.2 に示す A3～F5 の 5 音階を用い，対応する周波数へピッチシフト処理を行い声高を変更し生成した．複数の音階の利用には，特徴量の数を増やし機械学習モデルの精度向上を図る目的のほか，多音源音源と呼ばれる複数の音階で収録された音声ファイルを持つ UTAU 音源ライブラリへの対応が目的として存在する．多音階音源では歌声を合成する際，合成目標の声高に近い音階の音声を加工するため声高の変更による劣化を抑え，より自然な声を合成できる．特徴量抽出のためのピッチシフト処理においても，この多音階音源の仕様を考慮し適切な音声ファイルを選択し加工している．この音域ごとの音源ファイルの切り替えによる声質の変化を特徴量として捉え，多音階音源に対しより適切な評価スコアを推論できると考えた．

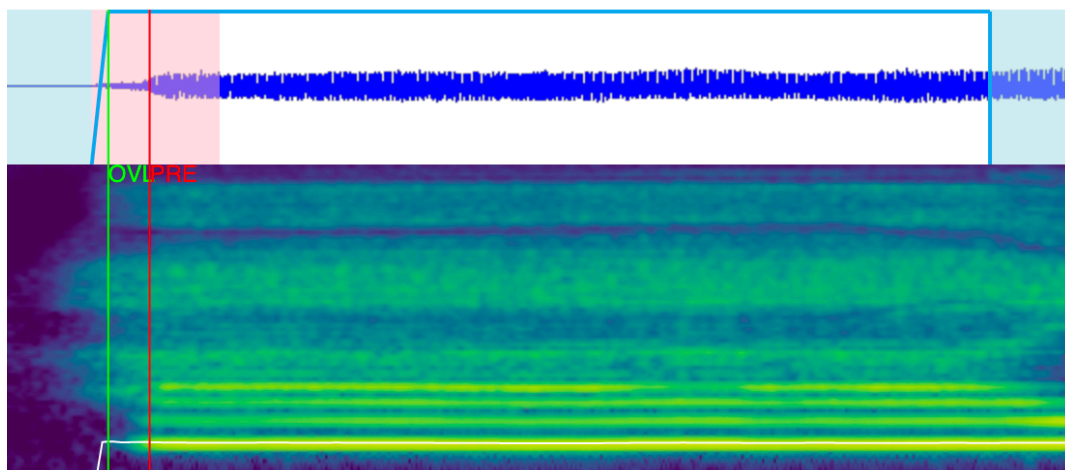


図 3.2: 音声ファイルの波形とスペクトログラム，ライブラリに指定されるタイミング

表 3.2: 使用する音階と Hz の対応表

音階	周波数 (Hz)
A3	220.000
D4	261.626
G4	391.995
C5	523.251
F5	698.456

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。Python ライブラリである librosa を用いて各音階と音素ごとにこれらの特徴量を抽出し、評価スコアの推論に用いた。librosa は音声信号処理のためのライブラリであり、音声ファイルを読み込み、スペクトログラムや MFCC などの音響特徴量を抽出する機能を提供している。今回設定した音響特徴量を全て抽出できるライブラリであり、また Python での利用が容易であるため採用した。表 3.3 に実際に抽出した特徴量の一部を示す。以下にそれぞれの特徴量についてとその採用理由について説明する。

表 3.3: 実際に抽出した特徴量の一部

音階	音素	MFCC_1	MFCC_2	MFCC_3	...	MFCC_64	ZCR	F1	F2	F3	F4
A3	a	-262.4385	69.4663	-77.1176		2.6309	0.0949	869.70	1432.72	3249.23	4601.21
	i	-321.4899	33.7375	-2.1761		3.5147	0.0831	274.35	2896.50	3616.97	4652.30
	u	-327.5042	60.8198	-29.5145		1.8560	0.0896	299.19	2705.92	3213.37	4281.72
	e	-229.1044	73.5700	-75.5765		3.3084	0.1362	559.06	2426.70	3371.12	4596.50
	o	-285.5089	115.6974	2.8371		-1.5311	0.0409	537.16	768.93	3727.63	4355.36
	N	-334.3087	81.4415	-44.6614		-0.6848	0.0389	257.92	2495.23	3245.69	4477.26
D4	a	-234.4008	14.7731	-88.2117		0.1172	0.1370	1148.51	1758.31	3851.63	4811.62
	i	-353.3810	-13.5760	-0.9987		8.6272	0.3198	506.51	3458.42	4325.31	5143.16
⋮											
F5	o	-358.9689	-40.6235	-42.1471		4.9630	0.1700	1344.94	2044.20	5015.81	5538.95
	N	-380.2289	-23.8642	-24.8103		-1.1690	0.1921	692.42	2947.32	4542.61	6567.64

MFCC はメル周波数ケプストラム係数の略で、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数列である。音声信号に対してフーリエ変換を適用して得られる周波数スペクトルを、人間の聴覚特性に合わせて変換（メルスケール変換）し、さらに対数変換と離散コサイン変換を適用して得られる。周波数スペクトルと比較して MFCC は聴覚特性に合わせた特徴量を持つため、人間からの聞こえ方の特徴をよく反映しており、小島らによる歌声の印象評価と音響特徴量の関係についての研究 [12] では MFCC が歌声の声質評価に有用だと示されている。声質の特徴に大きな影響があると考え、本研究では生成した音声から MFCC を 64 次元まで取得し特徴量として用いた。

ZCR は音声の波形が 0 を交差する頻度を表す指標であり、音声信号の時間軸上で波形が正から負、または負から正に変化する回数を示す。高い音声ほど周波数が高くなり波形の変化が多くなるため ZCR も高くなるほか、摩擦音のような雑音成分が多い音声でも ZCR は高くなる。音声の声を固定した今回の音声であれば、音声の透明感などに関連する指標として扱えると考え採用した。

フォルマントは周波数スペクトルのピークの周波数であり、周波数の低いものから順に F1, F2, といった具合に呼ばれる。フォルマントは声質や音素の音色に大きく関係する特徴量で特に母音の認識に重要とされており、音声の母音によってそのフォルマント周波数には一定の傾向がある。粕谷らの研究によれば、同じ母音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 [14], 声の性別感や年齢感に影響があると考えられる。それらの特徴の表現に期待し、本研究では F1~F4 のフォルマント周波数を特徴量として採用した。

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 は機械学習のワークフローを簡略化する AutoML ライブラリのひとつである。少ないコード量で機械学習のプロセスを実行でき、また様々な機械学習モデルを効率的に比較・評価し、最適なモデルを構築出来るためこのライブラリを採用した。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。

スコアの教師データには UTAU 音源声質アンケートの結果を用いた。UTAU 音源声質アンケートの結果は各ライブラリごとのニコニコ百科記事や、配布ページなどで掲載されているほか、それらを取りまとめたデータが Excel ファイルとして提供されており、本研究ではこのデータを利用した。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 ライブラリを学習対象とした。データセットは学習データとテストデータに分割し、ランダムに選択した 30% をテストデータ、残りの 70% を学習データとして使用した。

7つの評価軸それぞれについて PyCaret で選択できるモデルごとの R2 値を主な指標に推定精度を比較した結果、全体的に精度の優れていた AdaBoost Regressor を選択した。R2 値は決定係数とも呼ばれ、回帰分析の際に推論の精度を示す指標である。AdaBoost Regressor はアンサンブル学習の一種であり、複数の弱学習器を組み合わせることで強学習器を構築する手法である。弱学習機としては決定木を用いており、学習データの誤差を修正するために学習データの重みを調整しながら学習を行う。他のモデルでは Extra Trees Regressor が比較時高い精度を示したが、学習後の結果を見ると過学習の傾向が見られた。比較時の R2 値の高さに加え、学習後の結果を見ても過学習の傾向が見られなかったため、本研究ではこのモデルを採用した。

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する。テストデータは先述の通り学習データから各評価軸ごとに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。

図 3.3 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 RMSE (Root Mean Square Error: 二乗平均平方根誤差) を、エラーバーは誤差の標準偏差を示している。RMSE は実際の値と推測された値の差を二乗して平均を取った後に平方根を求めたもので、数値予測モデルの精度を評価する際によく用いられる指標である。結果を見ると、RMSE の最も小さい「滑舌」の軸では 0.72、最も大きい「声の性別」の軸では 1.27 を示しており、評価軸ごとに大きな差は見られなかった。RMSE は 0 に近いほど予測精度の高さを示し、評価スコアが 1 から 7 の数値を取る点を考慮すると、全体的に中程度の精度が得られたと言える。RMSE の平均は 0.98 となり、この結果は先行研究 [9] での報告値 0.96 に近く、音響特徴量を用いた機械学習モデルとして一定の妥当性を示した。

図 3.4 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。各軸ごとの相関係数を見ると、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 であった。この結果から、声質の評価スコアを音響特徴量から推測するモデルが一定の相関を持っていることが分かる。

7つの評価軸ごとの予測された値と実際の値との散布図を図 3.5 に示す。図を見ると、相関係数の高かった「声の強さ」は右上がりの傾向が見られ相関の存在が分かる一方で、最も低かった「声の年齢」では点が縦に長く分布している傾向が分かる。これは相関の弱さだけでなく、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現状を示している。このような傾向は程度の差はあるものの他の評価軸でも見られ、精度を下げて一因となっている。

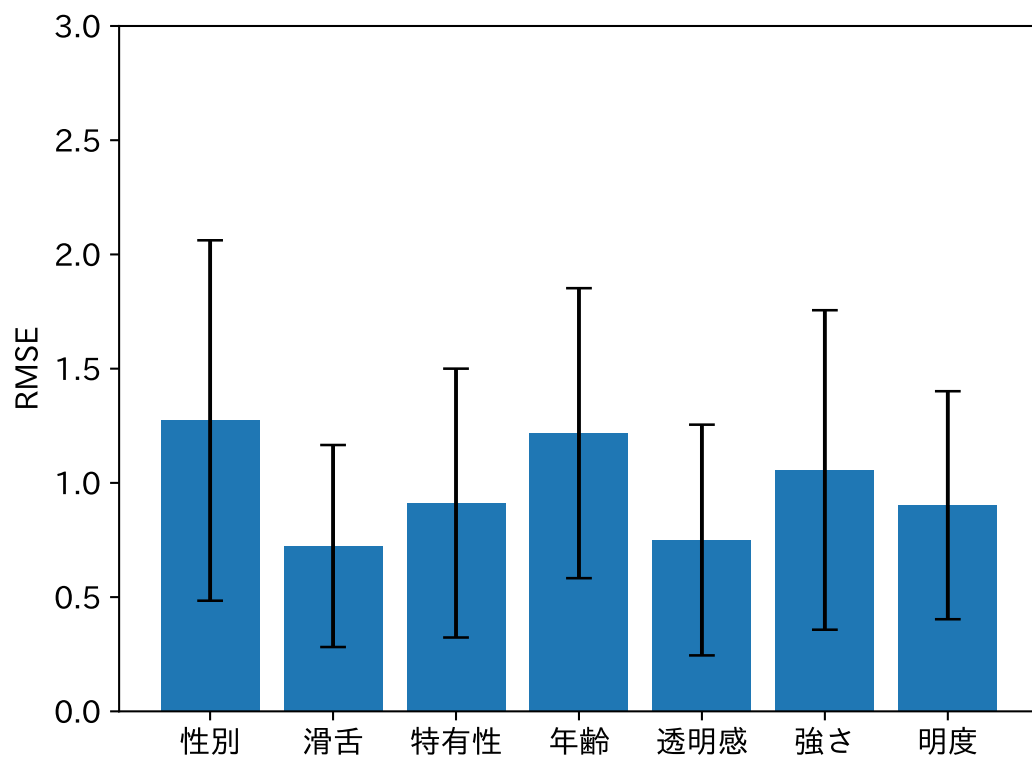


図 3.3: テストデータとの誤差の二乗平均平方根

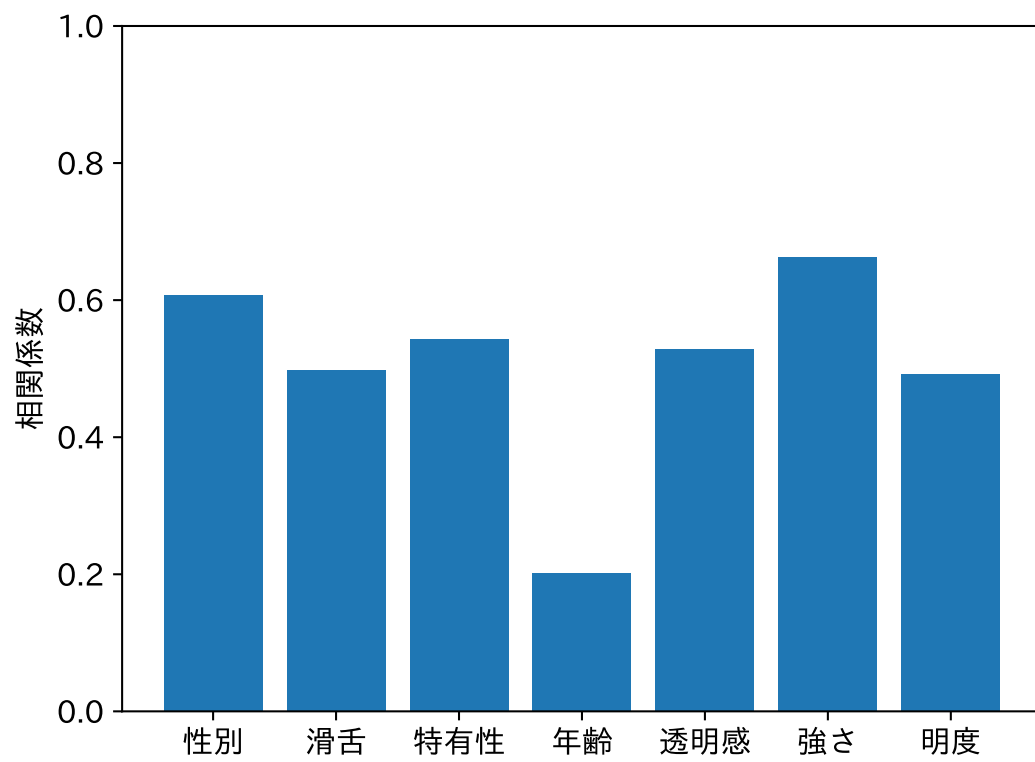


図 3.4: テストデータとの相関係数

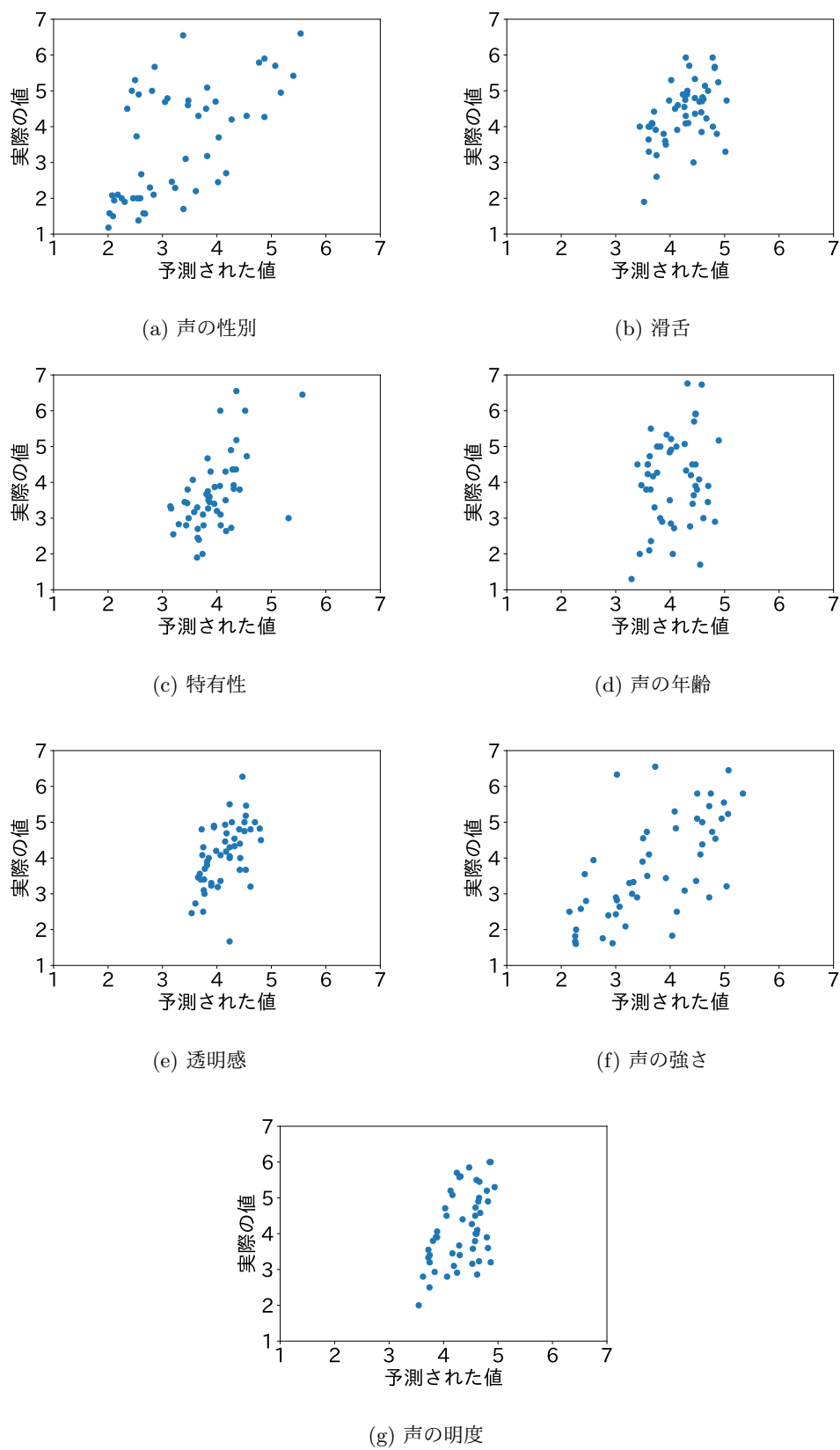


図 3.5: 実際の値と推測された値の散布図

第4章 声色見本帳

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。ユーザは自身が求める声に近いライブラリを、多くのライブラリの声を手探りで聞き比べる手間なく自身に合った声を探索できる。ライブラリの作成者も、自身のライブラリを多くのユーザに知ってもらう機会を得られる。実際に作成したサービスの画面を図 4.1 に示す。

声色見本帳

2.7

5

3

3

3

5

3

☒ 声の性別

☒ 滑舌

☒ 特有性

☐ 声の年齢

☐ 透明感

☒ 声の強さ

☐ 声の明度

男性的

はきはき

癖がある

大人びた

クリア

力強い

明るい

女性的

舌足らず

素直

幼い

ノイジー

優しい

暗い

<

1

2

3

4

5

...


24

>

検索

重音テト - 連続音

カサネテト



▶ 0:00 / 0:08

🔊

⋮

声の性別	滑舌	特有性	声の年齢	透明感	声の強さ	声の明度
2.67	5.33	3.27	3.6	4.47	5.2	5.4
-0.0	+0.3	+0.3	-	-	+0.2	-

波音リツ - 強連続音

ナミネリツ




図 4.1: サービスの画面イメージ

4.1 要求仕様

本研究で提案するサービスについての要求仕様を以下に示す。ユーザは求める声の評価スコアを数値として入力し、その評価スコアに近い UTAU 音源ライブラリを探索できる。探索結果として提示されたライブラリの歌声を実際に聞ける機能も必要である。評価スコアを用いて声質を表現できるとしても、実際に聞いてみて本当に入力した評価スコアが求める声に近いかを確認したり、探索結果として提示されたライブラリの中から好みの声を選べる。探索結果からは配布ページへのリンクを確認でき、ユーザがスムーズにライブラリをダウンロードできる環境を提供する。本サービスの想定するユーザの目的は求める声のダウンロードであるが、探索した後にライブラリをダウンロードする機能までは提供しない。これは、ユーザは本来ライブラリのダウンロード時にライセンスや利用規約を確認する必要があるが、それらのサービス内での提供が難しいためである。また、ユーザが新しいライブラリを追加できる機能も提供する。追加されたライブラリはサービスに登録されると同時に評価スコアが自動で推定され、以降の探索対象として利用される。この機能により、サービスに登録されていなかったライブラリや、今後公開されるライブラリも将来的に探索できる。サービスは Web サービスとして実装し、ユーザがブラウザから手軽に利用が可能な形式にする。合成音声ソフトウェアである UTAU はスマートフォンではなく PC での利用が一般的であるため、そのライブラリの探索も PC での利用が自然である。よってサービスは PC での利用を前提としてデザインを行う。

4.2 実装

サービスのフロントエンドは TypeScript を用いて開発し、フレームワークとして Next.js, UI コンポーネントライブラリとして Chakra UI を用いて実装した。バックエンドは Python を用いて開発し、フレームワークとして FastAPI を、データベースとしては PostgreSQL を用いて実装した。バックエンドを Python で実装したため、ライブラリを追加する際事前に構築した機械学習モデルによる推論やサンプル音声の合成を一挙に行える。

4.2.1 ライブラリ探索機能

探索はサービスのトップ画面から行う。ユーザは求める声の評価スコアを7つの評価軸に対してスライダーを用いて評価スコアを0.1刻みの小数で入力した後、探索ボタンをクリックして探索を行う。スライダーによる入力インタフェースに加えそのスライダーの両端に評価スコアの高低に対応する表現語を配置したため、数値である評価スコアと表現語の表す声との関係を直感的に理解できる。探索ボタンをクリックすると API を通じてデータベースにアクセスし、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。

入力の際、ユーザは各評価軸ごとに存在するチェックボックスを操作できる。チェックボックスはデフォルトでは全てチェックされており、チェックを外すと対応する評価軸を無効化できる。無効化された評価軸はスライダーもグレースアウト状態になり操作できなくなり、探索時にその評価軸が無視される。この機能は、ユーザにとって7つの評価軸全てに対する求めたい声の評価スコアの想定と入力は煩雑であると考え実装した。この機能によってユーザは自身の重視するいくつかの評価軸に対してのみ評価スコアを入力し、他の評価軸については入力せずに探索できる。

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を用いる。ライブラリに対して推定された評価スコアと、ユーザが入力した評価スコアとの差分を各評価軸ごとに計算した後、その差分の二乗和を計算して評価スコア間の距離を求める。この際、チェックボックスで無効化された評価軸については差分の計算を行わず、評価スコアの距離に影響を与えないようにする。距離が小さい順にライブラリを並び替え、探索結果として表示する。この方法は探索のたびに全てのライブラリに対して距離の計算を行うため、ライブラリの増加に伴い計算量が増加してしまうが、ライブラリ数が数百

程度であれば十分な速度で探索が行えたため、現状の規模では問題ないと判断した。今後登録ライブラリ数が増加し探索速度に大きな影響が見られた場合は探索の高速化を検討する必要がある。探索の高速化の方法としては、探索結果をキャッシュして再探索時の高速化を図る方法や、ライブラリをその評価値によって事前にクラスタリングし、クラスタごとに探索を行う方法などが考えられる。

また、探索結果として表示されたライブラリを選択し、その評価スコアを入力欄に転写できる機能を実装した。あるライブラリの評価スコアを探索にそのまま用いると、そのライブラリに近い声質のライブラリを探せる。

4.2.2 ライブラリ情報の表示

探索結果として、各ライブラリの名前やアイコン、推定された評価スコアなどライブラリの情報を表示する。評価スコアは評価軸ごとに探索に用いたスコアとの差分も表示し、求める声とどのような違いがあるかを一目でわかるようにした。ライブラリのスコア表示は探索時に入力したスライダーと同じ向きに同じ並びで表示されるため、余計なストレスなくスコアを比較できる。

ライブラリの声について実際に聞いて比較できるよう、ライブラリ情報の中にライブラリの歌声を再生できるメディアプレーヤを設置した。歌声の確認に用いるサンプル音声として事前に楽曲の一部を歌わせた音声ファイルを用いた。歌わせる楽曲としては童謡の「かえるのうた」を選定し、最初の一節を対象とした。

ライブラリ情報からライブラリの配布ページへ遷移できるリンクを設置した。ユーザはこのリンクからライブラリの配布ページにアクセスし、先方で示されるライセンスや利用規約を確認した上でダウンロードを行う。

4.2.3 探索対象ライブラリの追加

ユーザはライブラリ追加画面から未登録のライブラリを検索対象として登録できる。ライブラリ追加画面を図 4.2 に示す。追加する際はライブラリの名前やフリガナ、バージョン、配布ページの URL の情報を入力し、UTAU 音源ファイルを zip ファイルとしてアップロードする。バックエンドではアップロードされた zip ファイルを展開し、3 章で述べたように音響特徴量を抽出し、構築済みの機械学習モデルで評価スコアを推定する。UTAU 音源ファイルのアップロードには時間がかかるため、ユーザにアップロード進捗を伝えるためアップロード完了までプログレスバーを表示する。アップロード後はアップロードしたライブラリの詳細を確認できるページへ遷移する。登録処理もまた多少の時間がかかるため、サンプルボイスや評価スコアなどは遷移直後は表示されないが、推論と登録が終了し次第表示される。アップロードされたファイルは各種データの生成後には必要がなくなるため、データベースには保存せずに削除する。

アップロード時には合わせてサンプル音声の合成も行う。一般的に UTAU 音源ライブラリを扱うソフトウェアである UTAU や OpenUTAU は CLI 上での動作に対応しておらず、自動的な処理に対応していない。そこでサンプル音声の合成には、Python による音楽・歌声シンセサイザインタフェースである ScoreDraft を用いて合成を行った。このライブラリは Python から UTAU 音源ライブラリを用いた音声合成が行えるため、サーバ上での自動的な処理に適している。

これらの処理が完了すると、ライブラリの情報と推定された評価スコア、合成されたサンプル音声データベースに登録され、以降の探索で利用できるようになる。

声色見本帳

キャラクター名	<input type="text" value="歌音ナマエ"/>
フリガナ	<input type="text" value="ウタネナマエ"/>
バージョン名	<input type="text" value="通常連続音"/>
配布ページURL	<input type="text" value="https://example.com"/>
UTAU音源ファイル	<input type="text" value="zipファイルを選択"/>

登録

図 4.2: ライブラリ追加画面

第5章 おわりに

5.1 まとめ

本研究では合成音声ライブラリが多く存在する中で、ユーザが用途に合った声を探るための手法として、声の印象を数値化し、それを用いてライブラリを探索する Web サービス「声色見本帳」を提案した。まず、UTAU 音源の声から声の印象を数値化する機械学習モデルを構築した。モデルは入力データとして音声から抽出した MFCC をはじめとする音響特徴量を、出力データとして声の印象を 7 つの評価軸ごとに数値化した評価スコアを取る。入力データは Python ライブラリである librosa を用いて抽出し、学習時に用いる評価スコアには既存の UTAU 音源声質アンケートの結果を用いた。モデル構築には Python ライブラリである PyCaret を使い、学習方法には AdaBoost Regressor を、教師データとして声に対する既存のアンケート結果から 168 ライブラリ分のデータを用いた。構築したモデルの推論精度を確認するため評価実験を行い、評価軸ごとの相関係数が概ね 0.5~0.6 となる結果を得、一定の精度での推定を確認した。一方で、「声の年齢」をはじめとする複数の評価軸では予測値が評価スコア範囲中央に偏るなどの問題が見られた。

ユーザが求める声の印象を評価スコアとして入力し、声の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。ライブラリ探索に用いるライブラリごとの評価スコアは、先立って構築した機械学習モデルを用いて事前に推定し付与する。サービスはユーザの求める声に近いライブラリを複数提案し、それらのサンプル音声やダウンロードページへのリンクを提供する。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。

5.2 今後の課題

今後の課題としては、まず機械学習モデルの改善が挙げられる。改善手法として、より多角的な音響特徴量の利用が考えられる。本研究でモデルの入力として用いた音響特徴量は、主に母音の発声から特徴量を抽出している。しかし、子音にまつわる特徴量や、子音を含む音素間の遷移にまつわる特徴量などから声の印象を得られると考えられる。これらを用いてモデルの入力データの表現力を向上させれば、より適切に声の印象を推定できる。

また、運用されるサービスを活用し、学習データの拡充も考えられる。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と実際にダウンロードページにアクセスした音源との関連付けを行い、学習データとしての利用が考えられる。あるいはより直接的に、音源ごとにユーザから評価の投票を募り、その結果を用いるなど、ユーザからのフィードバックを利用した手法も考えられる。学習データをより多様で正確なものにすれば、モデルの汎用性向上が期待できる。

サービスとしての利便性向上も課題である。現在、ライブラリの新規追加時には UTAU 音源ファイルを zip ファイルでアップロードする形式であるが、このファイルは数百 MB のものも多く、アップロードに時間がかかる問題がある。サービスの運用にはこの音源ファイルそのものは必要ないため、ユーザのローカル環境上で音声ファイルの解析とサンプル音声の生成が出来れば、その結果を送るのみで済むため通信時間の大幅な削減が期待できる。探索システムの改善としては、現在評価スコアの類似度指標としてユークリッド距離を用いているが、それについて評価実験と検討が必要と感じている。より本当にユークリッド距離は評価スコアの類似度を適切に表現できるのか、また他の指標を用いた場合の探索結果の変化

について検討が必要である。また，ライブラリ情報の更新・修正機能，あるいは権利者からの削除依頼への適切な対応体制の整備も必須である。これらの改善を行いより利便性の高いサービスを目指す。

謝辞

本研究を進めるにあたり，多くの御指導，御鞭撻を賜りました梶克彦教授に深く感謝致します．
また，日頃から熱心に討論，助言してくださいました梶研究室のみなさんに深く感謝致します．

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] 田中 晶之, 中村 嘉志, ランダムフォレスト法を用いた歌声が聞き手に与える印象の単語推定に関する研究. 国土舘大学理工学部紀要, Vol.15, No.15, pp.23-28, 2022
- [5] 平山 健太郎, 伊藤 克亘, ポピュラー歌唱における高音域の声区と発声状態の判別手法. 情報処理学会研究報告, Vol.2012, No.16, pp.1-6, 2012
- [6] 鈴木 千文, 坂野 秀樹, 旭 健作, 森勢 将雅, 歌唱音声の類似度評価を目的とした基本周波数の変動量を反映するビブラート特徴量の提案. 電気学会論文誌C (電子・情報・システム部門誌), Vol.137, No.12, pp.1607-1614, 2017
- [7] 山本 雄也, 中野 倫靖, 後藤 真孝, 寺澤 洋子, ポピュラー音楽の模倣歌唱における歌唱テクニック分析と楽譜情報との対応付け. 情報処理学会論文誌, Vol.64, No.10, pp.1423-1437, 2023
- [8] , 山根 壮一, 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, ニュービッグ グラム, サクリ アニ サクティ, 中村 哲, 歌声合成システムの音源データ検索のための声質評価値推定法. 情報処理学会研究報告, Vol.2015-MUS-108, No.6, pp.1-6, 2015
- [9] 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, pp.1-6, 2015
- [10] 佐治 拓樹, 小林 和弘, 石黒 祥生, 戸田 智基, 大谷 健登, 西野 隆典, 武田 一哉, 声質の可視化を用いた所望音声検索システムの提案. 情報処理学会研究報告, Vol.2022-MUS-133, No.6, 1-5, 2022
- [11] 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う音楽発掘サービス「Ki-ite」を公開. https://www.aist.go.jp/aist_j/press_release/pr2019/pr20190830/pr20190830.html (2025 年 01 月 23 日閲覧)
- [12] 小島 俊, 齋藤 毅, 三好 正人, 歌声における印象評価と音響特徴量の関係について. 電子情報通信学会技術研究技術報告, Vol.111, No.471, pp.49-53, 2021
- [13] 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
- [14] 粕谷 英樹, 鈴木 久喜, 城戸 健一, 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, Vol.24, No.6, pp.355-364, 1968

- [15] 異式 連続音の録音リスト配布 - 異のブログ. <https://tatsu3.hateblo.jp/entry/ar426004> (2025 年 01 月 23 日閲覧)
- [16] 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> 音源声質アンケート (2025 年 01 月 23 日閲覧)