

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

K21066 清水洸世
指導教員 梶克彦

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とそのライブラリは非常に数多く存在しているが、それらを十分な数聞き込み目的に合った声を探し当てるには大きな手間がかかる。その問題に対し、声に対する印象を数値で表現する研究 [1] や、数値化した印象を利用し合成音声ライブラリの探索を行う研究 [2] は存在するが、ユーザが実際に利用できるようなサービスとしての提供を目的としたものは未だ無い。

本研究では、ライブラリの声に対する印象を機械学習によって自動的に数値化し、ユーザの求める声に近いライブラリを探索できる Web サービス「声色見本帳」を提案する。作成したサービス画面を図 1 に示す。探索対象は、特に利用できるライブラリが特に多く、後述する声質に関するアンケートが既存する UTAU 音源ライブラリとした。ユーザが手軽に探索できる実装に加えて、探索対象のライブラリの追加を自動的にを行いサービスを持続的に運営する仕組みを実装する。

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	-

図 1: 作成したサービス「声色見本帳」

2 評価スコアを推定するモデルの構築

本研究ではまず 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、ライブラリの音響特徴量から評価スコアを推定する機械学習モデルを作成する。音響特徴量として、ライブラリの音声から MFCC, ZCR,

F1~F4 の周波数を抽出する。評価スコアの教師データには、既存の UTAU 音源声質アンケートを用い、推定する評価の軸も同アンケートに基づく 7 軸を用いる。

本研究で利用する 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 [3] 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。

表 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

2.1 モデルの構築

本研究では、音響特徴量から評価スコアを推定するモデルとして AdaBoost Regressor を採用した。この選択は、評価スコアと音響特徴量とともに連続値である点、また予備実験において離散モデルでの学習時に過学習の傾向が見られた点に基づいている。

教師データには、アンケート結果をまとめた配布ファイルに記載されているライブラリの中から、学習に利用できる条件の整った 168 ライブラリを使用した。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。

モデルの構築には Python ライブラリ PyCaret を使用し、7 つの評価軸それぞれについて独立したモデルを作成した。各モデルは、先述の音響特徴量を入力として受け取り、対応する評価軸のスコアを出力として推定する。

2.2 モデルの評価

構築したモデルによる推論の精度を評価するため、テストデータに対する予測スコアと実際のスコアの相関係数を求め、その結果を表 2 に示した。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相関係数を示している。その結果、最も低い「声の年齢」は 0.20 とかなり低いものの、残りの 6 軸では 0.49 から 0.66 程度、そのうち最も高い「声の強さ」では 0.66

と、ある程度の相関が見られた。

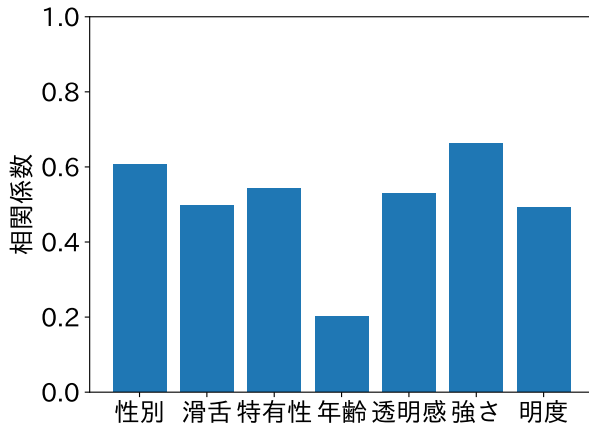


図 2: テストデータとの相関係数

相関係数が最も高かった「声の強さ」と最も低かった「声の年齢」について、実際の値と予測値の散布図を図 3 に示す。散布図を見ると、「声の強さ」では対角線上に沿った分布が見られるのに対し、「声の年齢」では予測値が中央値付近（3～5）に集中している。この予測値の中央集中傾向は、「声の性別」と「声の強さ」を除く他の評価軸においても同様に観察できる。このような傾向は、現在使用している音響特徴量では声質の印象を十分に表現できていないために生じていると考えられる。

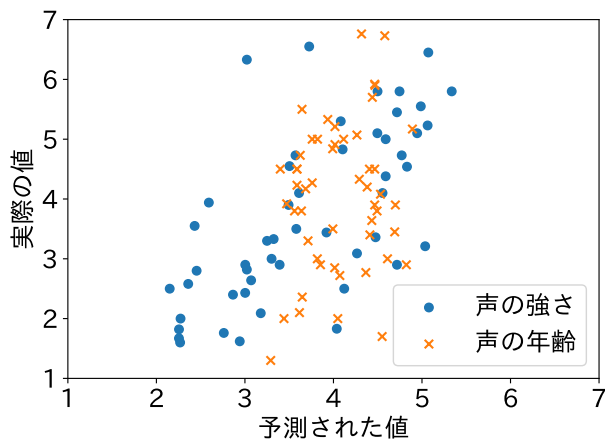


図 3: 実際の値と推測された値の散布図

3 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。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。

3.1 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サンプル音声を再生したり、配布ページへのリンクを確認できる。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。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。

3.2 実装

サービスはフロントエンドに React、バックエンドに FastAPI を用いて実装した。バックエンドを Python で実装することで、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うことができる。

ユーザはサービスのホーム画面から各評価軸に対してスライダーを用い 1～7 の評価スコアを入力できる。検索に用いる評価軸はチェックボックスを用いて選択でき、重視しない軸は無視して探索が行われる。入力された評価スコアとライブラリの評価スコアとの平方ユークリッド距離を計算し、距離が近い順に並び替え探索結果として表示する。

探索結果からは、各ライブラリの名前や評価スコア、アイコン、サンプル音声などを確認できる。サンプル音声には「かえるのうた」の一節を ScoreDraft を用いて事前に合成したものを使用する。また各ライブラリの詳細画面を開けば配布 URL などの情報を確認できる。

ユーザはライブラリ追加機能を用いて未登録のライブラリを検索対象として登録できる。追加する際はライブラリを zip ファイルとしてアップロードすると、サービスが自動的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。アップロード時には合わせて配布ページへの URL やライブラリの名前などのメタデータの入力を要求するほか、サンプル音声の合成も行われ、これらの情報もデータベースに登録される。

4 今後の課題

今後の課題を以下に示す。まずは 2.2 節でも述べたように機械学習モデルの改善が挙げられる。具体的な改善策としては、音響特徴量が声質の印象をより正確に表現できるよう子音の発音や音素間の時間的な遷移にまつわる特徴量を追加する案が考えられる。また、探索時の類似度指標の改善や、ライブラリ情報の修正・削除機能の実装などサービスの利便性向上も目指したい。

参考文献

- [1] 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
- [2] 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
- [3] 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