Vol.2022-MUS-133 No.6
2022/1/25

声質の可視化を用いた所望音声検索システムの提案

佐治 拓樹[1]　小林 和弘[2]　石黒 祥生[3]　戸田 智基[2]　大谷 健登[1]　西野 隆典[4]　武田 一哉[3]

概要：声質変換や歌声合成は，元となる音源データを入れ替えることにより，合成後の声質を容易に変更することができる．しかし，公開されている音声データセットは膨大なため，所望の声質のデータを探索することは非常に困難である．本研究では，声質の可視化を行い，その視覚情報を利用した声質検索システムの提案を行う．声質の可視化には，話者コードによって話者の特徴量を表現するサブネットワークを接続した VQVAE による声質変換モデルを使用する．サブネットワークで表現された各話者の特徴量を主成分分析によって二次元に削減することで，散布図の形で声質の可視化を行う．出力図の点をクリックすると対応した話者のサンプル音声を聞くことができるので，ユーザは可視化した図と実際に聞いた音声の関係を把握しながら所望の声質を探索していく．また，ユーザがシステムに探索したい声を声真似した音声を入力することで，出力図の改善を試みた．その結果，声真似を利用することで，データセットを可視化しただけの手法よりも探索の効率化を行えることがわかった．

キーワード：音声データ検索，検索システム，インタラクション

1. はじめに

楽曲の制作に使われる歌声合成技術や，望み通りの声に音声を変換できる声質変換技術が広く利用されている．歌声合成技術とは楽譜情報と言語情報を入力とすることで，任意の歌声を合成することができることができる技術である．またこの際，合成に使用される音源データや声質のパラメータを操作することで，合成音声の声質を変化させることも可能である．また，声質変換技術では所望の音声データを入力とすることでその声質へ自分の音声を変換することができる．しかし，これらの技術を利用するには，目標となる音声データが必要である．現在は多くの音声データセットが公開されているため，所望の音声データが存在する確率は高いが，膨大なデータの中からそれのみを探索するのは非常に困難である．よって，所望の音声データを抽出することができる音声検索システムの構築が求められる．そこで今回，音響モデルを利用して音声データセットの各話者の声質の特徴量を数値化した上で可視化し，その視覚情報を利用して所望の声質データを取得するシステムを提案する．また，その際ユーザが自分自身のイメージする声を声真似して入力することで，探索の効率化を試みる．

[1] 名古屋大学情報学研究科
[2] 名古屋大学 情報基盤センター
[3] 名古屋大学 未来社会創造機構, 株式会社ティアフォー
[4] 名城大学 都市情報学部

2. 関連研究

2.1 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

関連研究である音源データ検索システムを一つ紹介する．音声検索システムとして山根らは，主観的な声質評価値を指標とした声質評価値の自動推定法を提案している [1]．これは歌声合成システムの音源データを検索する際に，検索クエリに用いる指標として声質表現語を用いるために，音源データ一つ一つに声質評価値を付与する技術の構築を目的としたものである．具体的には，“年齢”，“綺麗さ”，“性別”，“滑舌”，“力強さ”，“癖の強さ”の六つの項目の評価値を音声データに自動で付与するモデルを構築し，検証を行った．検証の結果，“年齢”と “性別”に対しては各々 0.8 以上，0.9 以上の相関係数が得られており，高い推定精度．“滑舌”，“力強さ”は 0.6 程度，“癖の強さ”に対しては 0.4-0.6 程度の相関係数を得た．“綺麗さ”の相関係数は 0.1 以下であり，推定精度は低い物であった．しかし，この提案手法は検索の際，ユーザが所望する声質を各声質表現語で数値化する必要がある．それは困難と考えられるため，本研究では，ユーザが所望の音声を直接入力し，それをモデルが数値化するという手法を採用した．

2.2 話者コードによるモデル適応

本研究で必要となる声質の可視化のための音響モデル

に，話者コードを利用した音響モデルを使用する．Xue らは Deep Neural Network(DNN) の話者適応手法としてこのモデルを提案した [2]．これは話者コードと呼ばれる話者依存の離散値をサブネットワークを介して DNN に供給するモデルであり，各層のバイアス項を話者コードによって制御する．Xue らの手法では，あらかじめ通常の DNN に学習を行い，話者非依存のモデルを構築した後，話者の特徴を表現するサブネットワークを学習する．これによって話者依存の情報が，サブネットワークにある話者コードによる話者表現の重みパラメータに集約されることになる．今回は，この話者コードによって制御される話者表現のパラメータを利用して声質の可視化を行う．

3. 提案システム

3.1 提案システムの概要

音響モデルを使用して検索対象全ての音声データを可視化し，その視覚情報を元に音声を取りだす検索システムを作成した．提案システムを図 1, 2 に示す．図の各点は，音声データセット内の話者の声質の分布を表している．各点をクリックすると対応する話者のサンプル音声が流れるので，ユーザは分布の情報と声質の違いを確かめながら所望の音声を探索していく．

3.2 声質の可視化

提案システムで行っている声質の可視化についての説明を行う．音声データの声質の可視化は，小林らのオープンソースのモデル [3] を用いて行う．これは Variational Autoencoder(VAE) の一つである Vector Quantised-VAE(VQVAE)[4] をベースとした声質変換のモデルである．

使用モデルの VQVAE ネットワークのエンコーダーとデコーダには Aäron van den Oord らの Wavenet を元としたネットワーク構造を採用し [5]，ネットワークのデコーダ部分には，話者コードによって制御される話者特徴量を推定するサブネットワークが接続されている．VQVAE では本来，言語情報，音素，声の高さ，その話者の特徴などが全て潜在変数に表現されてしまうが，このサブネットワークを接続し，そこに F0 と話者コードによって制御される話者特徴量の行列を供給することで，潜在変数には言語情報や音素を，話者特徴量の行列にはその話者の特徴を表現するように学習することができる．特徴量は CheapTrick[6] で抽出されたスペクトル包絡線からパラメータ化されたメルフィルタバンクを用いる．またその複号のために ParalllelWaveGAN ボコーダを使用した [7]．このサブネットワークに表現された話者特徴量は一話者につき 32 次元の値で表現されているので，その値を主成分分析によって二次元空間に落とし込む．これをその話者の声質と捉えてシステムに出力した．学習には JVS corpus[8] のデータセットを使用する．話者数は男性 50 人，女性 50 人であり，それぞれの発話数は 100 である．

3.3 各手法の説明

今回提案する手法は三つである．点をクリックしてサンプル音声を聞きつつ，所望の音声を探索するという探索方法自体は各手法で変わらないが，出力されている図がそれぞれ違っている．

手法 1 は検索対象となる話者の声質の可視化を利用しただけのもので，手法 2，3 は声質の可視化に加えて，ユーザの入力した地声と所望の音声を声真似した音声を利用したものである．そのため，手法 2，3 では検索前にユーザが地声と所望している音声を声真似した音声をシステムに入力する．ここで言う「地声」はユーザが普段平素で話す「最も自然に出す声」として定義している．しかし，声真似というのは誰にでもできるものではないので，実際に所望している声質の音声とユーザがそれを真似して入力した音声には小さくない乖離があるはずである．そのため今回はユーザーの地声と声真似音声の特徴量の差を「ユーザーがイメージする声質へ向かうベクトル」と仮定し，手法を提案する．

各手法の説明は以下の通りである．

3.3.1 手法 1：音声データセットの可視化

手法 1 は図 1 に示した音声データの声質を可視化しただけの手法を表している．この手法では前述した話者特徴量を二次元に落とし込んだ値をそのまま出力している．

3.3.2 手法 2：手法 1 の軸を変更

手法 2 は手法 1 の軸を変化させたものである．手法 1 よりも所望の音声の探索に有効と思われる分布の図を出力した．前述した通り，ユーザに入力させた声真似の特徴量と地声の差の特徴量の差を欲しい声へ向かうベクトルと捉えて，その差分ベクトルの変化を捉えやすいような軸を設定し，出力している．具体的には x 軸をそれぞれの話者特徴量を差分ベクトルに射影した座標にし，Y 軸を差分ベクトルと直行した上で，声質のばらつきを最もとらえる方向とする．

3.3.3 手法 3：手法 1 ＋ユーザの地声と声真似の可視化

手法 3 は図 2 に示したように音声データの可視化に加えて，ユーザ自身の地声と声真似も可視化したものである．話者データの声質の分布は図 2 と同じものである．ユーザは声質の分布に加えて自身の入力した地声と声真似の位置を見て，所望の音声データを探索していく．

全ての図の赤色の点で表現された話者は，前述したモデルの学習話者であり，男性 50 人，女性 50 人の合計 100 人となっている．

4. 評価実験

4.1 実験方法

前述した提案手法三つを用い，実験的評価を行う．検索

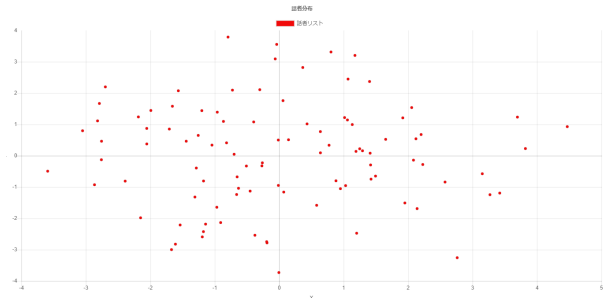

図 1　手法 1：音声データセットの可視化

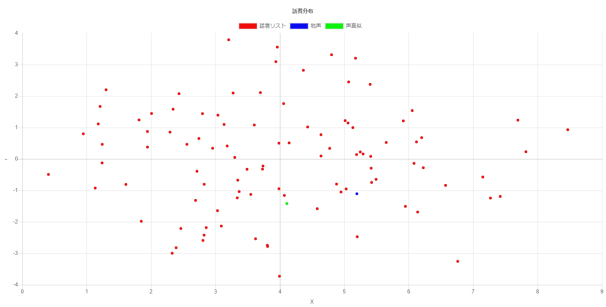

図 2　手法 3：手法 1 ＋ユーザの地声と声真似の可視化

タスクは実験協力者に特定の話者の声を聞かせた後，提案システムで同じ話者を検索するという形で設定した．本来であれば，ユーザが所望の音声をイメージしている状態で検索することを想定したシステムであるが，所望の音声がデータセットに存在しない可能性があること，正解のデータが存在した方が客観的な評価が行いやすいという二点を考え，この形をとった．

実験手順は以下の通りである．

(1) 実験協力者に検索対象である指定話者（100 話者からランダムに選択した 1 話者）の音声データを聞かせる
(2) 実験協力者は指定話者と同じ声をした話者を指定された手法（手法 1 － 3 のいずれかランダム）で検索する
(3) 実験協力者は納得いくまで検索をした後，指定話者の音声データと同じ話者と思った音声データの話者番号を選択
(4) 行った検索について主観評価を行う

実験協力者は検索中，指定話者の音声は何度でも聞けるようになっている．実験協力者に聞かせる指定話者の音声データは，学習話者の一人が指定のテキストを 20 秒ほどで読み上げている．読み上げるテキストはクリック時にシステムが流すサンプル音声のテキストとは異なったものとなっている．検索対象となる指定話者は実験協力者一人について 8 話者（男性 4，女性 4）である．よって，実験は 8 話者 × 3 手法の 24 通りの全てを完全にランダムな順番で上の手順で実施する．

評価は主観評価と客観評価どちらも行う．主観評価は検索のしやすさについての 5 段階評価（1: しにくい-5: しやすい）で検索が終わるごとに行う．客観評価は，探索時間

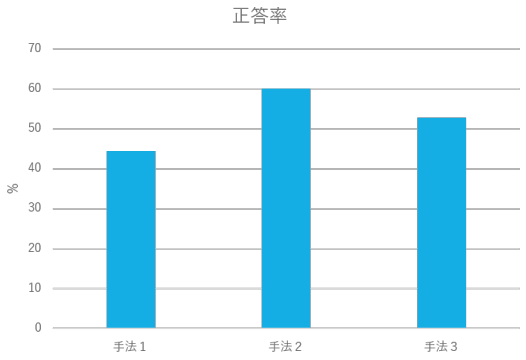

図 3　正答率

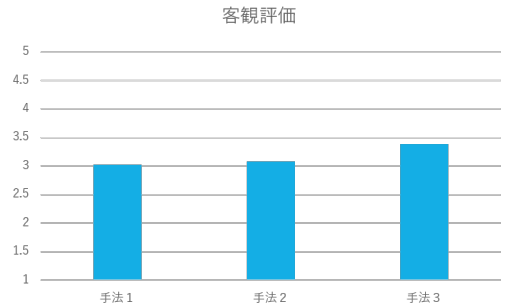

図 4　主観評価

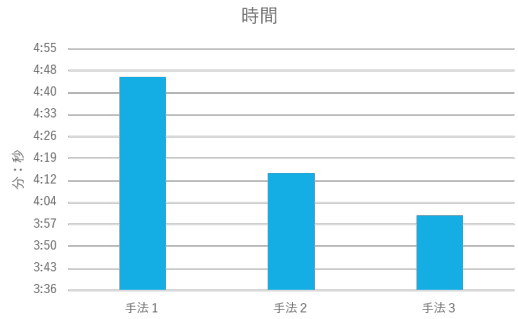

図 5　探索時間

（上の実験手順の 1 から 3 までの時間）と探索精度（実験協力者が指定話者を選べたかどうか）の二つについて行う．実験は 20 代男 6 名に対して実施した．

4.2　実験結果

正答率，主観評価，探索時間についての実験結果をそれぞれ図 3，4，5 にそれぞれ示す．正答率は実験協力者の選択した話者が指定話者と同じであった割合を表す．正答率に関しては，手法 2 が最も高い結果となった．次に主観評価については手法 3 が最も良い結果となった．各平均標準偏差は少数第三位を四捨五入したとき，手法 1 が 0.53，手法 2 が 1.65，手法 3 が 1.57 となった．最後に探索時間については，手法 1 が最も長く，手法 3 が最も短く 46 秒の差が存在した．また秒単位で各平均標準偏差を算出し，少数第三位を四捨五入すると，手法 1 が 170.38，手法 2 が

表 1 実験協力者の入力した声真似の平均順位（小数点第三位を四捨五入）

	男性話者	女性話者
平均順位	32.13	66.30
標準偏差	24.39	17.99

表 2 手法 3 における実験協力者の地声-声真似と地声-指定話者の角度（小数点第三位を四捨五入）

	男性話者	女性話者
平均角度	58.98	65.67
標準偏差	46.70	41.88

103.23，手法 3 が 92.74 という結果となった．

5. 考察

5.1 提案システムの使用方法について

提案システムの想定していた使用方法としては，図の「点の位置の変化」と「声質の変化」には，何らかの対応関係がある為，実験協力者は聞いた時にその変化を直感的に理解し，所望の声質のデータを探索するというものだった．手法 2 では軸を変化させることで，よりその対応関係を手法 1 よりもわかりやすいものに変えることで探索効率を上げようとし，手法 3 では，実験協力者自身の声質や声真似を入力することで，視覚的な検索の補助とすることで探索の効率化を行うことを試みた．

結果として正答率，客観評価，探索時間の三つの評価指標において手法 2，3 は手法 1 を上回ることができたが，実験協力者がとっていた探索方法は，想定していた探索方法とは違うものだった．まず例えば「右に行くほど年齢が高そうな声となっていった」というような一貫した変化を図の中に見出した実験協力者は存在しなかった．しかし，一貫した変化ではなく，似たような声質の話者がクラスターのように固まっているように感じたという意見が多く得られた．さらに手法 1，3 では男性話者が左下，女性話者が右下に固まっており，実験協力者も探索の内にそれを把握することが可能であった．そのため多くの実験協力者が，まず性別によって目星を付けた後，欲しい声質と似た話者を一人見つけ出し，その周辺を探索するという方法を取っていた．

よってそのクラスターごとに声質を言語化した何らかのラベルを付けることができれば，もしくはそのクラスターごとに区別をし，色分けなどを行って実験協力者に提示することができるだけでも探索の効率化が行えると考えられる．

5.2 声真似の精度について

今回の提案手法 2，3 では実験協力者が所望の声質を声真似した音声を用いた．しかし，実験協力者は特に声真似を得意としているわけではない一般男性であるため，その声真似の精度には不安が残る．特に検索対象となる話者が女性の場合，男性の実験協力者が異性である女性の声真似をするのは困難である．

そこでユーザの入力した声真似の特徴量が指定話者の特徴量と近い値を出すかどうかをモデルの話者推定機能によって確かめた．話者推定機能では入力された音声がモデルの学習した話者のどの話者かどうかを確率値で算出することができる．よって実験で得られた声真似データがどの話者に似ているのかを確率値で出力した後，推定確率が高いものから順位付けを行った．その後，男女それぞれで平均順位を算出した．

結果を表 1 に示す．結果を見ると，想定通り，男性話者と女性話者で大きな開きが存在した．手法 2，3 ではこの話者推定機能を元に声質の近い話者算出し，効率化を試みているので，表 1 を見ると，声真似による効率化を行うのは，同性の声を探索する場合に限ると考えられる．

5.3 ユーザの声のマッピングについて

手法 3 ではユーザが入力した地声と声真似を手法 1 の図にマッピングした．これはユーザの入力した地声と声真似のマッピングの先に所望の音声データが存在する仮定の元で提案した手法であった．そこで実験で取得したデータを用いてマッピングの先に検索対象の指定話者の音声データが存在するのかどうかを確かめた．

具体的には，まず実験協力者の地声から声真似と地声から指定話者の傾きを出した後，そこから二つの線分の角度を算出した．その後，男女それぞれで平均角度を算出した．

結果を表 2 に示す．結果を見ると，男性話者の平均角度は 58.98 度，女性話者 65.67 度とおよそ実験協力者の地声と声真似の線分を基準として 60 度以内に絞り込めることが判明した．

そのため，ズームして 60° 以内の話者のみを見せるような機能をシステムに追加することで，情報量の削減を図り，探索の効率化を行うことが考えられる．

6. まとめと今後の課題

今回は可視化を利用して所望の声質の話者の音声データを取り出す検索システムを提案した．さらにユーザが自身の地声と声真似を入力することで検索の効率化を試み，実験によって比較・検証を行った．結果として正答率，探索時間において手法 1 よりも手法 2，3 の方が良い結果を得られ，探索の効率化を行うことができた．

今後の課題としては，5 章で言及したような図の改善やユーザが探索しやすい機能の追加，そしてタスク条件を変えた実験の実施である．タスク条件を変えた実験とは，本来の目的により近い，検索対象の話者を探し出せたかどうかではなく，検索対象の話者と似た声質の話者を探し出せ

たかどうかというものを想定している．

謝辞 本研究の一部は JST，CREST，JPMJCR19A3 の支援を受けたものである．

参考文献

[1] 山根 壮一, 小林 和弘, 戸田 智基, 中野 倫靖, 後藤 真孝, Graham Neubig, Sakriani Sakti, 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, SIGMUS (2015).

[2] Shaofei Xue; Ossama Abdel-Hamid; Hui Jiang; Lirong Dai.: *Direct adaptation of hybrid dnn/hmm model for fast speaker adaptation in lvcsr based on speaker code*, ICASSP (2014).

[3] A toolkit for non-parallel voice conversion based on vector-quantized variational autoencode 入手先 ⟨https://github.com/k2kobayashi/crank.html⟩（参照 2021-12-20）.

[4] Van Den Oord, Aaron, and Oriol Vinyals.: *Neural discrete representation learning*, Advances in Neural Information Processing Systems (2017).

[5] Aaron van den Oord and Sander Dieleman and Heiga Zen and Karen Simonyan and Oriol Vinyals and Alex Graves and Nal Kalchbrenner and Andrew Senior and Koray Kavukcuoglu.: *WaveNet: A Generative Model for Raw Audio*, CoRR(2016).

[6] M. Morise: *Cheaptrick, a spectral envelope estimator for high-quality speech synthesis*, Speech Commun., vol. 67, pp. 1–7 (2015).

[7] R. Yamamoto and E. Song and J. Kim.: *Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram*, ICASSP (2020).

[8] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari.: *JVS corpus: free Japanese multi-speaker voice corpus*, University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 133–8656, Japan. (2019).