H31地域協働研究（ステージⅠ）

H31-Ⅰ-16「原敬の音声再現（原敬100回忌記念事業の一環）」

研究提案者：原敬100回忌記念事業実行委員会
研究代表者：ソフトウェア情報学部　榑松理樹

＜要旨＞

本研究では、原敬の音声再現の再現方法について、その手法を検討した。手法としては、大きく2つの観点に着目する。1つめは声質の再現である。ベースとなる人の顔写真の特徴点と原敬の顔写真の特徴点の差から声質変化量を決定する。このとき顔の特徴点の差と声質変化量の差は、音声･顔写真とも既知の人の情報を収集し、機械学習手法を適用することで求める。もう1つは話し方の再現である。文献などから、原敬の話し方を調べ、それをもとに声質変化量を決定する。これらの変化量をベースとなる人の声に適用することで声の再現を試みる。今年度は、基本設計を行ったが、実装には至らなかった。今後、実装を進めていく。

1　研究の概要（背景･目的等）

2020年は、第19代内閣総理大臣として活躍した盛岡出身の原敬の100回忌を迎える。平民宰相として岩手県民をはじめ国民に親しまれた原敬の顕彰と法要を行い、後世に意義深く継承するべく、2017年より原敬100回忌記念事業実行委員会を組織し、記念事業の計画、準備を進めている。

記念事業の目玉として計画しているのが、原敬の音声再現である。現在、原敬の音声は発見されておらず、原敬の肉声は誰も聞いたことがない。声優による再現は声優選定の根拠がなく、説得力に欠ける。一方で、現在分かっている原敬のデータ（骨格、体格、性格、話し方等）がある。これらのデータを活用することで、コンピュータによる原敬の音声再現を行うことが、本研究の課題である。

本研究においては、原敬の諸データを基にしたコンピュータによる音声を再現するための基本研究並びに合成音声の声質変換ソフトウェアの開発による音声再現、質問者の質問に再現した音声により政治姿勢や生き方等を回答する双方向のやり取りを可能にするソフトウェアの開発を目標として掲げる。本研究期間においては、主に、合成音声の声質変換ソフトウェアの開発による音声再現に主体を置き、基本的なアルゴリズムの確立およびプロトタイプの構築に取り組み、これらの実現を具体的な達成目標とする。また、声質変換ソフトウェアについては、ボイスチェンジャなどの類似ソフトウェアの知見が活用できる可能性が高い。そのほか統計的音声解析など、本研究を進めるうえで必要となる基本的なアルゴリズム･手法を活用する。

2　研究の内容（方法･経過等）

本研究で開発を目指すシステムの概念図を図1に示す。本システムは、読み上げる文章に対し、既存の音声合成ソフトウェアにより構築した音声を、独自に開発する声質変換ソフトウェアによって、声質（音声特徴）を変更することで、原敬氏の音声再現を目指す。声質変換ソフトウェアでの声質変換は、骨格（顔写真）の差と声質の差との関係に基づき構築する変換規則により行う。この情報は、複数の人の顔写真と音声データに対し、統計的音質変換と機械学習の手法を適用し、構築する。

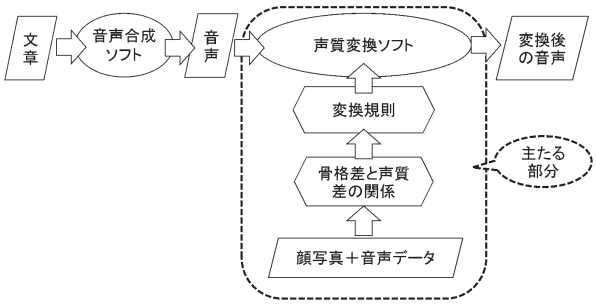

図1　概念図

本システムは、次のような考えに基づいている。

原敬の音声と合成音声があれば、統計的声質変換においては、入力話者＝合成音声、出力話者＝原敬の音声とすればよい。しかし、原敬の音声はないため、これは不可能である。

一方､原敬に関する情報としては､写真がある。その写真から口の大きさ（以後、骨格）など発声に影響を与える情報を一定範囲で得ることはできる。これに対し、合成音声話者については、写真はない。しかし、実際の人を介在することで、それぞれの部分を補うことが考えられる。よって、本研究では、実際の人を介在することで、合成音声の変換を試みる。具体的には以下の流れとなる。図１，２に概念図等を示す。

(1)　合成音声と介在者の声の違いから、合成音声話者と介在者の骨格の差を推定する。
(2)　推定した骨格の差を介在者に適応し、合成音声話者の骨格を推定する。
(3)　推定した骨格と原敬の骨格（写真）から、骨格の差を推定する。
(4)　推定した骨格の差から､声質の差（変化量）を求める。
(5)　合成音声に、声質の差（変化量）を適用し、音声の再現を試みる。

これらを行うには、骨格の差と声質の差との関係（骨格の変化に応じた音声特徴量の変化）を得る必要がある。こ

の関係を、複数人の声と写真のデータを収集し、それらに機械学習を適用することで求める。この際、声質の差を得るために統計的声質変換処理の考えを応用する。

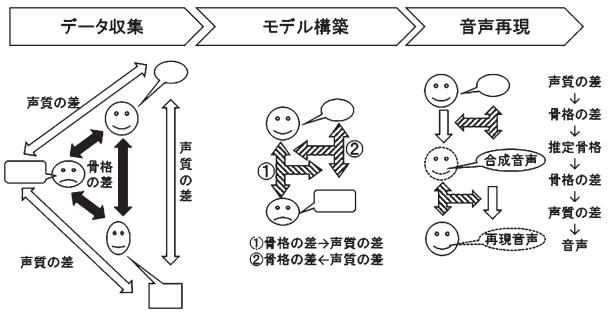

図２　研究の基本方針

３　これまで得られた研究の成果

今年度は２章で述べた考えをもとに、実現方法を検討し、図3に示す基本設計を構築した。以下、その内容を述べる。

基本的な方針としては、既存の声の音声特徴を変更することで、音声の再現を行う。2章で示したように、当初は合成音声を変更することを計画していたが、実在する人物の声を利用することに変更した。これは、合成音声の顔画像を予測することが困難であるためである。この点については、今後の課題とする。

音声特徴としては、基本的な特徴量であるピッチ、速度、音量に着目する。これらの音声特徴の変更量を決定する必要がある。そのために、2つの手法を検討した。1つめは顔画像を用いる手法であり、2つめは文献などを利用する方法である。前者の手法では、主にピッチの変化量を決定し、後者では、速度･音量の変化量を決定する。以下、それぞれについて説明する。

はじめに、声質の再現方法について述べる。顔画像を用いる手法においては、ベースとなる人の顔画像と音声のほか、複数名の顔画像と音声を収集する。これらに対し、ベースとなる人の顔画像を中心と、顔画像の差分をとる。基本的には輪郭、口の位置、大きさなどの発声に関連するところのみに注目する。具体的には、顔画像に対し、一定のルールに基づき12か所の点を付与する。そのうちの中央に位置する点を基準点として、各点をその点から見た距離と角度を用いて表記する。この時、各自の頭頂とあごの先端との距離を1とし、各点との距離は、この距離を基準とした相対距離で示す。また合わせて音声解析によって得られる特徴量、主としてピッチの差を得る。ピッチとしては、母音を対象とし、発話区間を対象とし、F0（基本周波数）および第1フォルマントから第4フォルマントまでの値を抽出する。これらを学習データとし、機械学習手法を適用することで、骨格の違いと周波数の違いの関係規則（モデル）を構築する。このモデルを使い、ベースとなる人とターゲットとなる人の画像の特徴点の差からピッチの変化量を決定する。

次に文献を用いる手法について述べる。本手法では、文献中におけるターゲットとなる人物の声の特徴、話し方の特徴を抽出する。主として、話し方（話速、大きさ、抑揚など）に着目する。それらの語彙をもとに、声質変化量を決定する。たとえば、ゆっくりとした話し方であれば、話速を遅くするという形である。これについては、演劇などを中心とした経験則をベースとし、変化させた音声と主観的な調査から調整することで実現を目指す。

上記双方の手法で得られた変化量に基づいて、ベースとなる人の声を変化させることでターゲット人物の音声の再現を試みる。

今回の音声再現には正解がないため、検証においては、実在の人物の声をターゲット音声とし、評価を進めることを検討している。

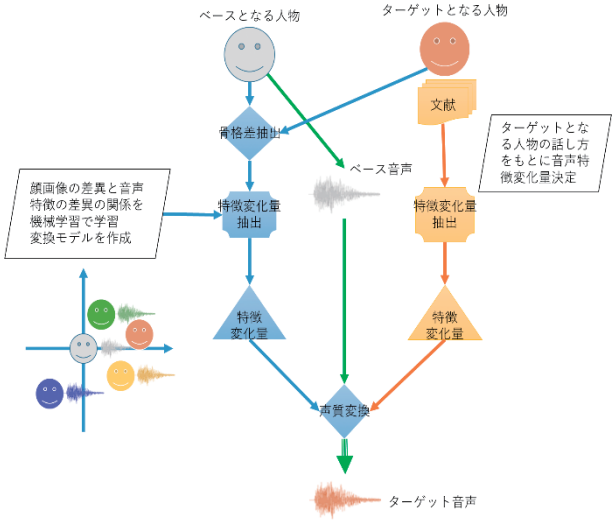

図3　基本設計

今年度は基本方針を固めるとともに、いくつかのモジュール設計･開発にとどまっている。また文献については、共同研究者より提供をうけ、現在調査中である。

これらの点から、研究の進捗状況は不十分と言わざるを得ない。

４　今後の具体的な展開

今年度は３章で述べたように基本設計構築にとどまっている。今後は、各種ツールなどを使い、システムを実装するとともに、必要な情報の収集を進める。顔画像および音声としては、オープンソースのものを利用する予定であるが、研究参加者を募ることや、ニュース映像などの利用も検討している。

正解がない目標ではあるが、記念行事の一環として、有用な成果を出せるよう実装を進めていく。

５　その他（参考文献･謝辞等）

研究実施にあたり、文献を提供していただいた原敬記念館館長･山内昭氏に感謝いたします。