多層ニューラルネットワークを用いた声質変換アルゴリズムの提案

Voice Conversion Algorithm Using Deep Neural Networks

佐久間洋司　中村泰

Hiroshi Sakuma, Yutaka Nakamura

大阪大学
Osaka University

1. はじめに

話し合いや他者とのコミュニケーションにおいて声質変換が重要な意味を帯びる場合がある．そもそも発話された音声には，韻律情報に加え話者情報や感情情報が含まれるが，例えば構音障害者が参加する話し合いや，複数の参加者が同時に発話するような場合において，声質変換によりコミュニケーションが改善される可能性がある．話し合いを円滑に進める手法の一つとして，話者変換に限らない，多様な声質変換を実現することが考えられる．

近年，人間の神経回路の構造を模した多層ニューラルネットワークを用いた，深層学習と呼ばれる手法が画像認識や音声認識など幅広い分野で注目を集めている．声質変換の研究においても線形写像による変換を前提とする GMM に代わって応用が試みられつつある[1]．

ここでは，深層学習により画像の画風を変換するアルゴリズム[2]が実現されていることに着目し，同様の手法で音声を変換することができないか検討する．具体的には，発話された音声を元に計算されるスペクトログラムの 2 次元的なマップを，画風を変換する場合と同様に変換することで話者性だけを変換した音声を生成できないか試みる．

2. スペクトログラム

スペクトログラムは声紋鑑定や音楽，音声処理などに用いられている 2 次元的なマップであり，一般的には横軸で時間を表し縦軸で周波数を表す．各点の明るさや色によってその点での周波数の振幅すなわち強さを表す．つまり，ある短時間の音声波形データにどのような周波数成分が含まれているかを示している．スペクトログラムを求めるには離散フーリエ変換を計算する必要があるが，複素関数 f(x) の離散フーリエ変換 F(t) は以下で定義される．

$$\mathrm{F(t)} = \sum_{x=0}^{N-1} f(x) e^{-i\frac{2\pi t x}{N}}$$

Python と SciPy ライブラリの関数群を用い図 1 のように容易に計算できる．

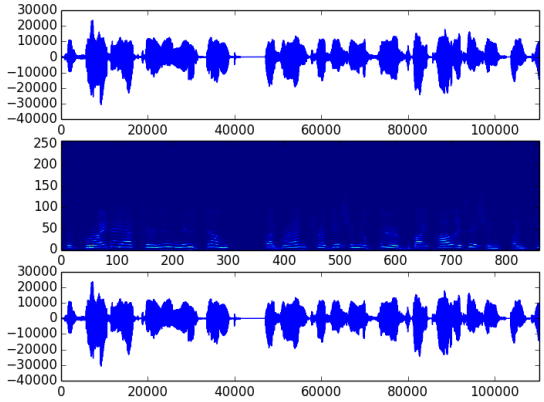

図 1 元の音声（上）と FFT で変換したスペクトログラム（中），逆変換して復元した音声（下）

3. 画風を変換するアルゴリズム

画風を変換するアルゴリズムは，画像認識用に学習されたネットワークを用いて，コンテンツ画像として入力された画像の物体の配置を維持したまま，画風をスタイル画像として与えられる画風に変換した画像を生成する．例えば物体認識用に学習させた 16 層の CNN (Convolutional Neural Network) のひとつである VGG 16-layer を用いる場合，ネットワークは図 2 のように表せる．

入力画像の 3×256×256 と書かれているうち先頭の 3 がチャネルにあたる．入力層におけるチャネル

には RGB の色情報が含まれ，また，画像の縦横それぞれ 256 ピクセルの情報が含まれている．図の通り，層が深くなるほどチャネルが増え，重要な特徴量が相対的に強くなるように学習されていく．その中で，情報が弱まっている部分に細工することで画風を変換することができる．

このそれぞれの中間層のチャネル間の相関を計算したスタイル行列という表現を導入する．入力層においては Red, Green, Blue の相関などが表現され，深い中間層ではどのような色と色が隣り合って存在しやすいかなどが表現されていることが示唆されている[2]．スタイル画像として与えられた画像に似せたスタイル行列で，コンテンツ画像のスタイル行列を差し替えることで, 画風を変換することができる．具体的には，中間層のコンテンツ画像との差分とスタイル行列のスタイル画像との差分を目的関数として最適化を行う[3].

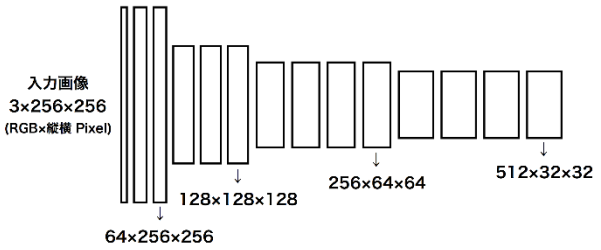

図 2 VGG 16-layer CNN モデルの構造

4. 声質変換アルゴリズムの提案

声質変換アルゴリズムを構築するにあたって，音声をスペクトログラムで扱うことで，前述の画風を変換するアルゴリズムで用いたような画像認識ネットワークを模した音声認識ネットワークを構築することを提案する．目標は音声認識ネットワークを用いて声質変換を行うことである．音声認識ネットワークは前述の画像認識のための多層ニューラルネットワークに倣って構築することができるため，上手く動くことが期待される．

今回は話者のラベルがついた音声のデータセットが与えられ，入力された音声を 0.5 秒程重複させながら 数秒単位に分割し，FFT で変換することでスペクトログラムを得る．このスペクトログラムを入力，ラベルを出力としてニューラルネットワークの学習を行う．

その学習済みの音声認識ネットワークについて，スタイル音声とコンテンツ音声を受け取り，それらの 2 つの音声のスペクトログラムを求める．前節で紹介したように，スタイル音声として与えられた音声のスペクトログラムと似せたスタイル行列で，コンテンツ音声のスペクトログラムのスタイル行列を差し替えることで，音声の声質を変換することができるはずである．その後，得られた変換済みのコンテンツ音声を IFFT で逆変換することで，声質が変換された音声を得ることができる．

5. 実験

今回の実験では，Apple 製コンピュータに付属する音声読み上げシステムを say コマンドによって利用し，音声データを取得した．say -v Agnes -o Agnes-0001.aiff -f news.txt のように実行すれば news.txt で用意したテキストを Agnes-0001.aiff で出力する．Agnes, Alex, Bruce, Fred の 4 システムによる約 4 分半のニュースの読み上げを行った．

なお，人間の可聴域は 20～20,000Hz と言われているが, 電話回線の周波数は 300～3,400Hz であり[4], 今回は後者の周波数を再現できれば十分と考えられる．って，音声の長さを 5.20s 取ることにし，ウィンドウ幅を 0.1s, スライド幅を 0.02s 取り，0.1s 幅の細かな音声が 256 個取る．縦幅である周波数は，直流である 0 から 5120 Hz 程度をカバーすれば良く，直流側から 512 次元分を持ってくれば，256 × 256 のサイズでスペクトログラムが用意でき，画像認識で使ったネットワークと同様の 2 次元配列のサイズになる．

学習結果の識別率をきちんと検討した上で，変換を試み，女性の読み上げ文を男性風に，逆の場合も声質変換する，このシステムについて知らない被験者に音声を聞かせて反応を調査し，後にどのような音声を被験者に聞かせたかを説明する. それにより，この声質変換アルゴリズムに対する正解率が示され妥当性が検証できる．研究会ではその結果を紹介するとともに，同システムがコミュニケーションを改善する可能性について検討する．

参考文献

[1] 中鹿, 亘. (2015). 深層学習に基づく声質変換(画像・音声・音声認識・理解,<特集>人工知能分野における博士論文) / Voice Conversion Based on Deep Learning. 人工知能:人工知能学会誌 / Journal Of The Japanese Society For Artificial Intelligence, (1), 131.

[2] Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A Neural Algorithm of Artistic Style.

[3] 画風を変換するアルゴリズム| Preferred Research https://research.preferred.jp/2015/09/chainer-gogh/

[4] 聴覚. Wikipedia, https://ja.wikipedia.org/w/index.php?title=%E8%81%B4%E8%A6%9A&oldid=56388712