50148

あなたの声を真似る人工知能

ARBMの複素拡張に基づくボコーダー不要な非パラレル声質変換

電気通信大学情報理工学研究科　中鹿 亘

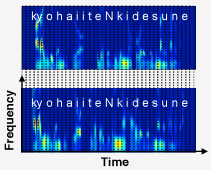
背景

声質変換とは

話している内容はそのままに，音声を特定の人の声に変換する技術。福祉やエンターテイメント用途の他に，喉頭摘出者のコミュニケーション補助等に使われる。

パラレル型と非パラレル型

モデルの学習にパラレルデータ（同一文言による音声データ対）を用いるアプローチを**パラレル声質変換**，任意の音声データで学習を行うアプローチを**非パラレル声質変換**と呼ぶ。パラレル声質変換は比較的高品質な変換が可能であるが，使用する学習データが限られてしまい，利便性が損なわれる。一方，非パラレル声質変換は任意発話で学習できるため利便性・実用性は高いが，品質面で課題がある。



パラレルデータの例

従来手法：ARBMによる声質変換

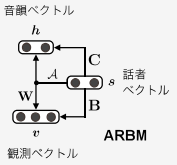
[T. Nakashika *et al.*, 2016]

ARBM: Adaptive restricted Boltzmann machine

アプローチ

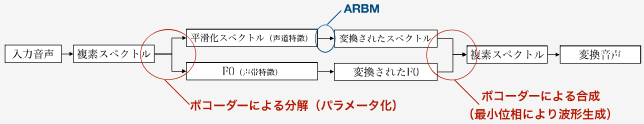
適応型制限ボルツマンマシン（ARBM）を用いた**非パラレル声質変換**手法。観測ベクトル，潜在的な音韻，話者を表すワンホットベクトルの間に存在する接続重みを自動学習させる。結果，観測ベクトルが与えられた時，音韻と話者を推定でき，話者を切替えることで声質変換を実現。

エネルギー関数に基づく生成モデル（確率分布）　$p(\boldsymbol{v},\boldsymbol{s},\boldsymbol{h}) = \frac{1}{U}e^{-E(\boldsymbol{v},\boldsymbol{s},\boldsymbol{h};\mathbf{W},\mathbf{B},\mathbf{C},\mathcal{A})}$，　$E(\boldsymbol{v},\boldsymbol{s},\boldsymbol{h}) = -\boldsymbol{v}^\top(\mathcal{A}\circ\boldsymbol{s})\mathbf{W}\boldsymbol{h} - \boldsymbol{h}^\top\mathbf{C}\boldsymbol{s} - \boldsymbol{v}^\top\mathbf{B}\boldsymbol{s}$



考えられる問題点

ボコーダーを使用する必要があり，最小位相の仮定やF0推定エラーで品質が低下してしまう。

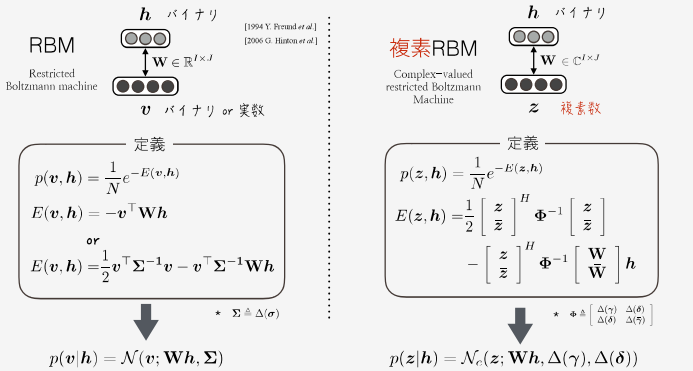


複素RBMの検証

[T. Nakashika *et al.*, 2017, 2018]

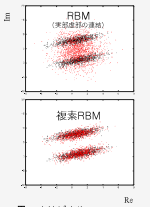
提案手法を検証する前に，まず，一般化したモデル（RBMを複素拡張したモデル；複素RBM）について評価し，実験的に有効性を確認した。

複素RBMと，通常のRBMの違い



人工データを用いた実験

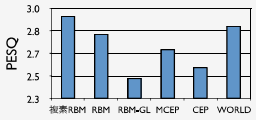
人工的に作成した１次元の複素数データを用いてRBMと複素RBMを学習させた。右図より，複素RBMは実部・虚部間の相関を捉えており，より元の**複素数データ分布を表現**できていることが分かる。



黒：オリジナル
赤：モデルから生成した乱数

音声データを用いた実験

200文の音声データを用いて複素RBMを学習させたところ，従来のRBMだけでなくメルケプストラム，ケプストラム，WORLDなど**代表的な音声特徴表現手法を上回る**結果が得られた。



提案手法：複素ARBMによる声質変換

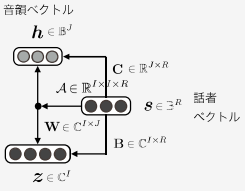
従来ではボコーダーを使用する必要があり，品質低下を招いていた。そこでボコーダーを使用せず，複素拡張モデルによって**複素スペクトルを直接変換**する手法を提案する。

$$p(\boldsymbol{z},\boldsymbol{h}|\boldsymbol{s}) = \frac{1}{Z}e^{-E(\boldsymbol{z},\boldsymbol{h}|\boldsymbol{s})}$$

$$E(\boldsymbol{z},\boldsymbol{h}|\boldsymbol{s}) = \frac{1}{2}\begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix}^H\boldsymbol{\Phi}^{-1}\begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix} - \begin{bmatrix}\boldsymbol{b}+\mathbf{B}\boldsymbol{s}\\ \bar{\boldsymbol{b}}+\bar{\mathbf{B}}\boldsymbol{s}\end{bmatrix}^H\boldsymbol{\Phi}^{-1}\begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix} - 2(\boldsymbol{c}+\mathbf{C}\boldsymbol{s})^\top\boldsymbol{h} - \begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix}^H\boldsymbol{\Phi}^{-1}\begin{bmatrix}(\mathcal{A}\circ\boldsymbol{s})\mathbf{W}\\ (\mathcal{A}\circ\boldsymbol{s})\bar{\mathbf{W}}\end{bmatrix}\boldsymbol{h}$$

$\star\ \boldsymbol{\Phi} \triangleq \begin{bmatrix}\Delta(\gamma) & \Delta(\delta)\\ \Delta(\delta) & \Delta(\gamma)\end{bmatrix}$

複素ARBM
Complex-valued adaptive restricted Boltzmann Machine



導出
①入力話者から音韻の推定　$p(\boldsymbol{h}|\boldsymbol{z},\boldsymbol{s}) = \mathcal{B}(\boldsymbol{h};\sigma(2\Re(\mathbf{W}^H(\mathcal{A}\circ\boldsymbol{s})^\top\bar{\boldsymbol{z}}) + 2\boldsymbol{c} + 2\mathbf{C}\boldsymbol{s}))$
②目標話者のスペクトル推定　$p(\boldsymbol{z}|\boldsymbol{h},\boldsymbol{s}) \quad \mathcal{N}_c(\boldsymbol{z};(\mathcal{A}\circ\boldsymbol{s})\mathbf{W}\boldsymbol{h}+\boldsymbol{b}+\mathbf{B}\boldsymbol{s},\Delta(\gamma),\Delta(\delta))$
を用いて声質変換

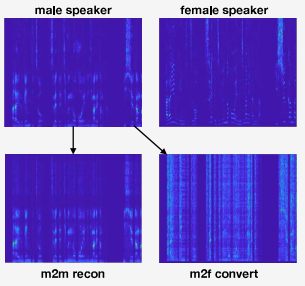
評価実験と現状の課題

男女１名それぞれ50文の音声を用いて複素ARBMを学習させ，実際に声質変換を実行したところ，右図のようなスペクトルが得られた。**同一話者のリコンストラクションは正しく行われた**一方，話者の特徴を表すフォルマントやF0など，**変換音声では正しく表現されていない**ことが分かる。この主な原因として，

課題1. 複素スペクトルをテンプレートの線形変換で表現していることが不十分

課題2. 同じテキストでも潜在変数（音韻）が異なる

などが挙げられる。



・可視素子=129次の複素スペクトル
・隠れ素子=2000次

$$p(\boldsymbol{z}|\boldsymbol{h},\boldsymbol{s}) = \mathcal{N}_c(\boldsymbol{z};(\mathcal{A}\circ\boldsymbol{s})\mathbf{W}\boldsymbol{h}+\boldsymbol{b}+\mathbf{B}\boldsymbol{s},\Delta(\gamma),\Delta(\delta))$$

複素スペクトルテンプレートWを行列Aで射影

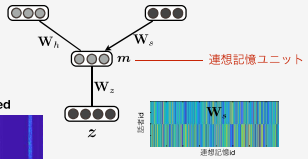
課題1. の解決策　連想記憶モデル

複素ARBMではスペクトルがテンプレートの線形変換で表現される仮定をしているため，F0やフォルマントを正しく表現できなかった。そこで複素ARBMに連想記憶を追加し，**複素スペクトルを非線形的に表現**できるモデル（連想記憶モデル）について検討した。

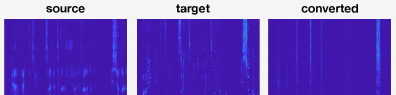
複素スペクトルの条件付き確率

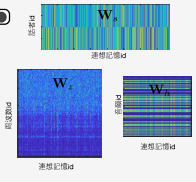
$$p(\boldsymbol{z}|\boldsymbol{m}) = \mathcal{N}_c(\boldsymbol{z};\underline{\mathbf{W}_z\boldsymbol{m}(\boldsymbol{z},\boldsymbol{h},\boldsymbol{s})} + \boldsymbol{b}_z,\Delta(\gamma),\Delta(\delta))$$

どのテンプレートWを使用するかをmで選択



声質変換例





連想記憶による非線形変換化により，**多少品質が改善**された。推定されたWsは一方が0他方が1となる重みが多く，mがそれぞれの話者に関する複素スペクトルテンプレートを選択することを示唆している。一方Whでは，各音韻素子に大きな差異が見られない。

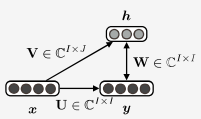
課題2. の解決策　複素CRBMによるパラレル声質変換

複素ARBMでは潜在的な音韻情報の共通化・差別化が困難であるため，パラレルデータを用いて**音韻情報を強制的に一致**させた場合の複素スペクトルの声質変換について検討した。CRBMの複素拡張（複素CRBM）を用いてモデル化し，パラレル声質変換実験を行った。

CRBM: Conditional restricted Boltzmann machine

複素CRBM

Complex-valued conditional restricted Boltzmann Machine



$\boldsymbol{x}$：入力話者の複素スペクトル
$\boldsymbol{y}$：目標話者の複素スペクトル
パラレルデータ

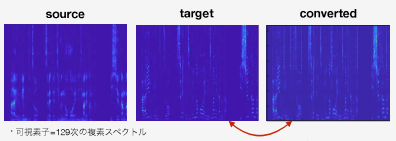
モデル学習後は，yの初期値を適当に設定し，

① $p(\boldsymbol{h}|\boldsymbol{x},\boldsymbol{y}) = \mathcal{B}(\boldsymbol{h};\sigma(2\Re(\mathbf{V}^H\bar{\boldsymbol{x}} + \mathbf{W}^H\bar{\boldsymbol{y}}) + 2\boldsymbol{c}))$

② $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h}) = \mathcal{N}_c(\boldsymbol{y};\mathbf{U}\boldsymbol{x} + \mathbf{W}\boldsymbol{h} + \boldsymbol{b},\Delta(\gamma),\Delta(\delta))$

を繰り返して目標話者の複素スペクトルyを推定

パラレルデータを使用することで**品質が著しく改善**し，真の目標話者に近い音声が得られた。また，線型変換と，潜在変数yを介した非線形変換の組み合わせによる変換も品質改善に貢献。複素ARBMでも何らからの方法で潜在音韻を一致させることができれば十分な声質変換性能を期待できる。



・可視素子=129次の複素スペクトル
・隠れ素子=2000次
・1時刻前のスペクトルも条件として加えている