調音可視化に向けた深層学習による音声からの口腔形状推定

情報科学科　梅村 直人　　指導教員：入部 百合絵

1 はじめに

日本人の英語の発音学習にあたり, 近年ではコンピュータ上で音声認識を利用したCAPT(Computer-Assisted Pronunciation Training)システムの導入が盛んに行われている[1][2]. しかし, CAPTシステムは学習者の誤った発音を矯正するための具体的な方法は示してくれない.

本研究では, 学習者が発音した際の調音器官の動作を視覚的に分かりやすく表現するため, 調音器官の動作を学習者の音声から推定することを目的とする. ネイティブによる正しい発音動作も示すことで, 学習者は可視化された自分の発音と模範となる発音を比較することで, どの調音器官をどのように動かせばよいのか容易に理解できる.

2 提案手法

提案手法の概要を図 1 に示す.

本研究では音声から発音動作を推定するために, 学習者の音声から声道情報を示す音響情報を抽出する. 音響情報には, 音声認識の特徴量としてよく用いられるmfcc(メル周波数ケプストラム係数)を採用する. 声道情報を抽出できれば, 発音動作に関連する情報を得ることができる. そして, その音響情報を入力とし, 調音器官（口唇, 舌, 口蓋垂など）の輪郭を示す座標値を出力とする音声－座標変換モデルを構築する. 時々刻々と変化する調音器官の位置や形をこれらの座標値によって表すことができ, 発音動作アニメーションの生成も可能となる.

3 使用するデータと深層学習

従来の発音動作の学習支援には模範となる口の動作を顔の正面から捉えた動画などが多く使われてきた. しかし, これでは舌や口蓋などの調音器官の動きや位置についての詳細な情報を得ることができない. 本研究では, 人体に電磁波を当てて断層撮影をするMRI動画像(図2)に基づいて, 調音器官の位置情報として座標値を抽出する. 座標値の抽出にはLucas-Kanade法を用いる. MRI動画像は口腔系内を側面から撮影しているため, 調音器官の動きを詳細に

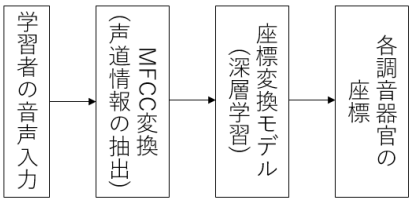

図 1 提案内容の流れ

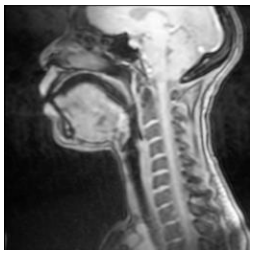

図 2 MRI動画像

表 1 各ネットワークにおける実験環境

	LSTM	CNN
学習/評価データ	102/8	102/8
Epoch 数	100	100
バッチサイズ	8	8
中間層	4 層	2 層
活性化関数	PReLU	Sigmoid
損失関数	HuberLoss	HuberLoss

表 2 調音器官毎の相関係数

	LSTM	CNN
上唇	0.777	0.801
下唇	0.425	0.516
舌先	0.635	0.697
口蓋垂	0.451	0.418
舌盛上	0.331	0.297
平均	0.524	0.546

捉えることができる.

音響情報から調音器官の座標へ変換するモデルの実現には深層学習を活用する. 深層学習には, 時系列データを扱うことに優れているLSTM(Long short-term memory) と, 画像に用いられるCNN(Convolution Neural Network)を利用した.

4 評価実験

抽出した音響情報と座標情報をもとに, LSTM とCNN をそれぞれ用いて学習と評価を行った. 実験環境を表1に示す. 正解座標と推定座標の相関係数を調音器官毎に平均した結果を表2に示す.

各調音点の相関係数を見ると, 上唇および舌先に関しては比較的高い相関係数であったが, 口蓋垂や舌の盛り上がり部分に関しては0.5を下回ることが多かった. これらの器官は動作の変化が大きいため, 精度よく座標に変換することができなかったと考えられる. そのため, 変換モデルに使用する特徴量の見直しが必要である. また, 学習データ数の不足, モデル構築方法や調音点の追跡精度の不十分さも原因として挙げられる.

5 まとめ

本研究では, 学習者が発音する際の調音器官の動作を視覚的に分かりやすく表現するため, 調音器官の動作を音声から推定する手法を提案した. 実験結果より, 正解値と推定値との相関係数が約 0.55 程度であり, 全体的に低い結果となった. 調音器官によっては音声から動作情報を得やすいものと得にくいものがあるため, 特徴量の見直しやモデル設計の改良などにより, 精度向上を目指す.

参考文献

[1] 河合他, 日本音響学会誌,Vol.57,No.9,pp.569-580(2001)

[2] Maxine Eskenazi, Speech Communication, Vol.51, No.10, pp.832-844,(2009)