6ZE-03

発音動作可視化を目的とした口腔形状の特徴量抽出

梅村　直人† 入部　百合絵†
愛知県立大学 情報科学部‡

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][3]．CAPTシステムは学習者が発話した音声に対し，音素ごとに誤り箇所を指摘，あるいは正しい音声波形と学習者の誤った音声波形とを比較表示する．これにより，学習者は自身の発音の誤りに気付くことができる．しかし，実際にどう口を動かせばいいかといった具体的な方法は理解し辛いものとなっている．したがって，発音の学習支援システムはどの調音器官をどのように矯正すればよいか示すべきである．

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ようになる．

2 音声からの発音動作推定手法

2.1 提案手法の概要

提案手法の概要を図 1 に示す．

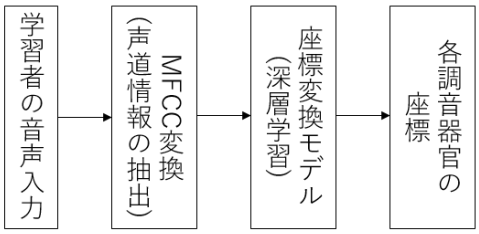

図 1　提案内容の流れ

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．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，最終的には座標値を利用した発音動作アニメーションの生成といった表現が可能である．変換モデルに用いたデータや音響情報および座標の抽出方法については以降に詳述する．

Estimation of articulatory motion from speech based on LSTM for articulatory visualization
Naoto UMEMURA, Yurie IRIBE†
Aichi Prefectural University, School of Information Science and Technology†

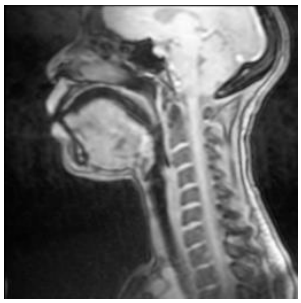

図 2　使用した MRI 動画像

2.2 調音器官の動作推定に使用した MRI 動画像

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 動画像に基づいて，それらの調音器官の調音様式や位置を視覚化する（図 2）．MRI 動画像は口腔系内を側面から撮影しているため，正面からビデオ撮影した映像では不明瞭であった調音器官の動きを詳細に捉えることができる．英語ネイティブ 2 名（女性）が英単語を発音する様子を撮影し，作成された MRI 動画像は英単語計 110 語分である．MRI 動画像には撮影時の英語ネイティブの音声も含まれる．

2.3 音響情報の抽出

撮像した MRI 動画像の音声には撮影時に発生した機械音やサイン音が混じっている．そのため，音声から音響情報を正確に抽出するためノイズフィルタを用いた雑音除去を施した．音声から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)を抽出する．mfcc はノイズ除去後の音声からフレーム毎に 12 次元を抽出した．mfcc には声道に関する情報が含まれているため，調音器官の動作に関連した特徴を得ることができると考えられる．

2.4 調音器官の座標抽出

変換モデルの出力データである座標はフレーム毎の調音器官の輪郭点を示している．複数ある調音器官の中でも，発音動作に大きく係わる上唇，下唇，舌先，口蓋垂，舌の盛り上がり部

分の 5 つの調音器官を表現することにした．各調音点の座標を正確に取得するため， Lucas-Kanade 法によりフレーム毎の調音点の追跡を行った．この手法では動画像の 1 フレーム目に該当する調音点を手動で指定すると，2 フレーム以降は自動で該当する調音点が追跡され，フレーム毎の座標値を抽出することができる．また，MRI 動画像は発話者ごとに顔の位置がずれているため，座標値は絶対値を採用するのではなく，各調音器官の 1 フレーム目を原点とした相対座標を求めた．さらに出力値は-1 から 1 の範囲に収める必要があるため，先述した 5 つの調音器官の座標値から最大値を求め，フレーム毎の座標値を最大値で除した正規化処理を実施した．

2.5 音声−座標 変換モデルの構築

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．深層学習の学習データおよび評価データを作成するにあたり，MRI 動画像からフレーム毎の口腔形状の画像と音声データをそれぞれ抽出した．先述したように画像からはフレーム毎に調音器官の輪郭座標を，音声からは mfcc12 次元を得る．そして mfcc を入力データ，調音器官の座標を出力する，音声ー座標値変換モデルを構築する．構築したモデルは各調音器官の座標値をフレーム毎（17ms/frame）に出力する．音声から精度よく座標値を出力するために，変換モデルには，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．

3 学習と評価結果

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．深層学習には SONY の Neural Network Console を用い，実験を行った．

学習曲線を図 3 に示す．また，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．図 3 より学習誤差（learning error）は 0 に収束しているのにも関わらず，テスト誤差(validation error)は上昇しており，過学習が発生している可能性がある．また表 2 より，

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

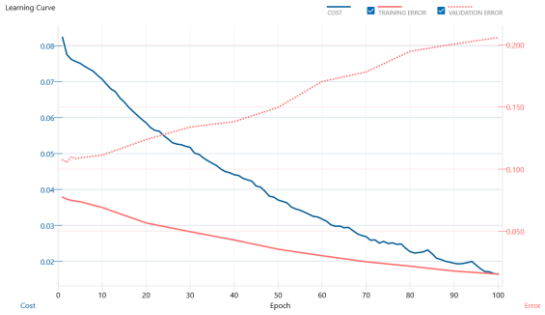

図 3 学習曲線

表 2 各調音器官における相関係数

	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

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，精度よく座標値に変換することができなかったと考えられる．そのため調音器官から取得する特徴点の位置の見直しや，特徴点を補う情報として調音器官の声道面積も取得するなど，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．

4 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．

謝辞

本研究は JSPS 科研費 JP15K00487 の助成を受けたものです．

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)
[3] S. Wang, M. Higgins, and Y. Shima, “Training English pronunciation for Japanese learners of English online,” *The JALT Call Journal, 1*(1), 39–47, (2005)