言語特徴量から調音運動の予測*

☆脇田真子，△髙畑諒汰，竹本浩典（千葉工大），平井啓之，前川喜久雄（国語研）

1　はじめに

音声の生成過程は，肺から送り出された呼気流を音源波に変換する発声と，音源波に言語としての音響的特性（音韻性）を付与する調音からなる。調音は舌，口唇，下顎，軟口蓋といった調音器官の運動によって声道の形状を制御すること(調音運動)で実現される。したがって，調音運動の情報を抽出して音声認識や声質変換などへ応用する研究が多数行われている[1]。

また，言語特徴量から調音運動を予測する研究も行われている[2]。この論文では，音素の種類やフレーム位置情報などの言語特徴量から，EMA（Electro Magnetic Articulography）で得られたセンサコイルの位置情報を予測している。しかし，センサコイルの個数が少数であるため，調音器官全体の形状を予測することは困難である。

そこで本研究では，まず調音器官全体の形状を観測できるリアルタイム MRI（rtMRI）を用いて調音運動を動画として記録し，フレームごとに調音器官の輪郭を抽出した。そして, DNN を用いて言語特徴量からフレームごとの調音器官の輪郭点（調音運動）の予測を試みたので報告する。

2　材料と方法

2.1 MRI 実験・録音

実験参加者は日本人成人男性 1 名（60 代）である。実験参加者は MRI 内で仰臥し，ATR503 文[3]を 2 日に分けて読み上げた。その調音運動を rtMRI で動画として記録し，音声も録音した。撮像は ATR-Promotions に設置されている Siemens 製 MAGNETOM Prisma fit 3 で行い，フレームレート 27 毎秒，解像度 1×1 mm，スライス厚 10 mm，画像サイズ 256 × 256 pixel であった。撮像中の音声は Optoacoustics 製 FORMI タイプ光マイクロホンを用いて録音した。

2.2 データセット

2.1 で得られたデータから実験で用いるためのデータセットの作成を行う。

まず，発話音声データに音素ラベリングを行い，得られたラベルデータに基づいて言語特徴量を作成した。これは，HTS[4]のフルコンテキストラベルから生成される当該音素や隣接音素の種類，アクセント情報などの一般的な DNN 音声合成用の特徴量パラメータに継続時間長や F0 を加えた 292 次元のベクトルで，動画のフレームに対応する時間ごとに計算される。

次に, 2 段階モデル抽出法[5]を用いて rtMRI 動画の各フレームから Fig. 1 で示す調音器官の 5 部位の輪郭を pixel 座標の点群として自動抽出した。なお，輪郭点は舌 40 点，口唇・下顎 40 点，軟・硬口蓋 30 点，咽頭後壁・披裂部 28 点，喉頭蓋・声帯 30 点であり，168（輪郭点数）×2（各点の x，y 座標）次元のベクトルである。

以上の処理で得られた言語特徴量と輪郭点を合わせてデータセットとし，1 日目の撮像で得られた 259 の発話のうち，239 の発話を訓練データ，10 の発話を検証データ，10 の発話をテストデータとした。

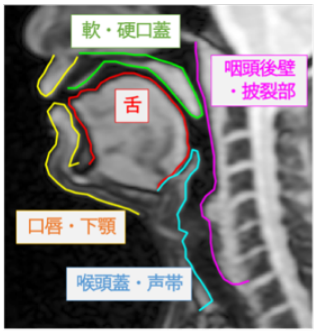

Fig. 1 調音器官の 5 部位

2.3 調音運動予測モデル

本研究では，全結合層を用いて言語特徴量から調音運動（フレームごとの調音器官の輪郭点）を予測するモデルを提案する。Fig. 2

* Prediction of articulatory movements from linguistic features, by WAKITA, Mako, TAKAHATA, Ryota, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (NINJAL).

に示すように，このモデルは入力として 1 発話の言語特徴量データを受け取り，2 つの全結合層を通して調音器官の輪郭点を出力する。各中間層次元数は 1024，出力次元数は各部位の輪郭点数×2 の次元数とした。損失関数には平均二乗誤差，最適化手法には学習率 0.01 の確率的勾配降下法を用いた。

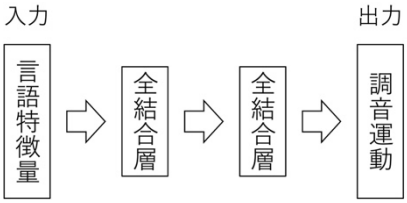

Fig. 2 調音運動予測モデルの構成

2.4 予測精度の評価

本モデルを用いて言語特徴量から予測したテストデータの輪郭点と正解（ground truth）となる自動抽出により得られた輪郭点との二乗平均平方根誤差（RMSE）を求め，精度を評価した。

3 結果と考察

Table1 はテストデータから発話の前後や文の間のポーズなど調音運動が休止する区間のフレームを除いた全 1085 フレームの各部位で算出した誤差である。予測した輪郭点と正解点との平均 RMSE は 1.19 pixel であった。

Fig. 3 は/a/と/i/を当該音素とするフレームのうち，平均的な予測精度をもつフレームにおける予測した輪郭線（輪郭点を結んだ線）と正解の輪郭線の比較で，両者は目視では良く一致した。Table 2 は部位ごとの平均 RMSE で，0.84 から 1.62 の間に分布した。

Takemoto *et al.* [6]は，MRI 画像から機械学習を用いた調音器官の輪郭抽出において，抽出した輪郭点から正解の輪郭線までの最短距離（垂直距離）の平均が 1 pixel 以内であれば，良好な結果であると評価した。本研究では，予測した輪郭と正解の輪郭の対応する点同士の RMSE を求めているため，予測した輪郭点が正解の輪郭線上にあっても誤差となる。そのため，Takemoto *et al.* [6]の評価に比べて誤差を過大評価していると考えられる。これを踏まえると，本研究では言語特徴量から調音運動を高い精度で予測できたといえる。

Table 1 各部位の予測精度（単位：pixel）

部位	x 座標	y 座標	平均
舌	2.11	1.76	1.94
口唇・下顎	0.87	1.00	0.94
軟・硬口蓋	0.63	0.75	0.69
咽頭後壁	0.67	0.99	0.83
喉頭蓋・声帯	1.41	1.63	1.52
平均	1.14	1.23	1.19

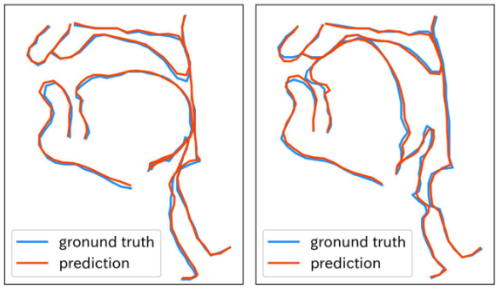

Fig. 3 予測結果と正解の比較（左: /a/，右: /i/）

Table 2 Fig. 3 の各部位の予測精度（単位：pixel）

当該音素	/ a /	/ i /
舌	1.01	1.30
口唇・下顎	1.45	1.62
軟・硬口蓋	1.39	0.84
咽頭後壁	0.99	1.05
喉頭蓋・声帯	1.19	1.24

4 まとめ

本研究では，DNN を用いて言語特徴量から調音運動の予測を試みた。予測と正解との平均 RMSE は 1.19 pixel であった。この数値は Takemoto *et al.* [6]の結果と直接比較できないが，目視では予測は良好であった。今後は，調音運動から音声信号を予測し，調音運動の予測誤差が音声信号の予測にどのような影響を与えるかを吟味し，これに基づいて予測精度の基準を設ける必要がある。

謝辞

本研究は JSPS 科研費 20H01265 の助成を受けて実施した。

参考文献

[1] K. Richmond *et al.*，音響学会誌，71 (10), 539-545, 2015.
[2] Z. Wei *et al.*, Proc.APSIPA ASC 2016, 1-6, 2016.
[3] 小林ら, 音響学会誌, 48 (12),888-893,1992.
[4] H. working group, 2021,"HMM/DNN-based Speech Synthesis System [HTS]," http://hts.sp.nitech.ac.jp/. (参照 2023-07-11).
[5] 藤澤ら，音講論 (秋)，1015-1016, 2022.
[6] H. Takemoto *et al.*, Proc.Interspeech2019, 904-908, 2019.