音声信号から発話器官の輪郭形状の逆推定*

☆梶浦一真，△関根諒，脇田真子，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

音声の生成過程とは逆に音声から調音運動を推定（逆推定）する研究が行われている。これは Acoustic to Articulatory Inversion（AAI）として知られ，音声合成，音声認識，発話トレーニングなど様々な音声処理分野への応用が期待されている。先行研究では声道形状のデータと音声データを同時に計測したコーパスに基づくものなど，様々な手法で逆推定が試みられている。その多くは EMA（Electro Magnetic Articulography）で得られた位置データを用いている[1]。しかし，EMA ではセンサコイルの個数や位置に制限があるため，調音器官全体の形状を推定することが困難である。

そこで本研究では，まず調音器官全体の形状と運動を観測できるリアルタイム MRI（rtMRI）を用いて調音運動を記録し，フレームごとに調音器官の輪郭点（調音運動）の推定を試みたので報告する。

2　材料と方法

2.1 rtMRI 撮像・録音

本研究では，60 代日本人成人男性 1 名による ATR503 文の発話を ATR-Promotions に設置されている Siemens 製 MAGNETOM Prisma fit 3rtMRI で撮像し，同時に音声も収録した。これは脇田ら[2]と同一のデータで，フレームレート 27 fps，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。

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を作成した。rtMRI 撮像と同時に収録した音声からスペクトルサブトラクション法[3]を用いて MRI 装置のノイズを除去した。そして WORLD（音声分析合成システム）[4]を用いて，ノイズ除去後の音声から 1 次元の基本周波数（F0），45 次元のスペクトル包絡（Mel Generalized Cepstrum），2 次元の非周期性指標（Aperiodicity）の合計 48 次元のベクトルを音響特徴量として抽出した。

次に，２段階モデル抽出法[5]を用いて rtMRI の動画の各フレームから調音器官５部位の輪郭を pixel 座標の点群として半自動抽出した。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）の合計 336 次元のベクトルである。

以上の処理で得られた音響特徴量と輪郭点を合わせてデータセットとし，重複した発話を含む合計 550 発話のうち 440 発話を学習データ，55 発話を検証データ，20 発話をテストデータとした。

2.3 調音運動推定モデル

本研究では，音響特徴量から調音運動（各フレームの調音器官の輪郭点）を推定する後述の２種類のモデルを比較した。推定は調音器官の部位毎に行った。また, 入力には, rtMRI の撮像フレームの中心時刻と，その前後-40, -20, -10, -5, 0, 5, 10, 20, 40 ms における合計 432 次元の音響特徴量を用いた。ただし，フレーム周期は WORLD が 5 ms，rtMRI が 37 ms であるため，線形補間により rtMRI のフレームの中心時刻に合わせた音響特徴量を用いた。

2.3.1 全結合型ネットワーク（DNN）

１つ目のモデルは 7 つの隠れ層をもつ全結合層のニューラルネットワークで，活性化関数には ReLU を用いた。隠れ層のユニット数は[2048, 1024, 512, 256, 512, 1024, 2048]とした。

2.3.2 畳み込みネットワーク（CNN）

２つ目のモデルは，間に２層の Self-Attention 層を含んだ６層の逆畳み込み層を持つニューラルネットワークである。入力音響特徴量は全結合層により 512 チャンネル×１に変換し，

---

* Inverse estimation of speech organ contours from speech signals, by KAJIURA, Kazuma, SEKINE, Ryo, WAKITA, Mako, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (National Institute for Japanese Language and Linguistics).

その後調音器官の輪郭形状のサイズまでアップサンプリングすることにより推定する。活性化関数には ReLU を用いた。舌の場合の各層のチャンネルは[4096, 2048, 1024, 512, 256, 128, 1]である。発話器官毎にポイント数が異なるため，stride, padding を調整し必要な出力サイズになるようにした。

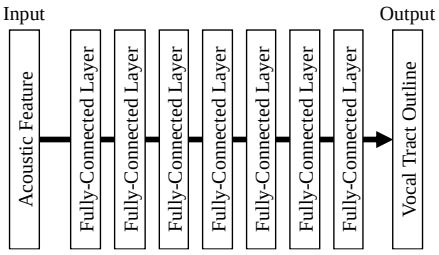

Fig. 1 全結合ネットワークの構成

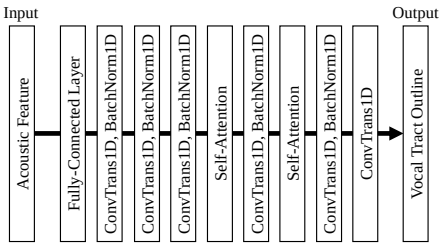

Fig. 2 畳み込みネットワークの構成

2.4 予測精度の評価

各モデルを用いて音響特徴量から推定したテストデータの輪郭点と正解（ground truth）となる半自動抽出により得られた輪郭点との平均絶対誤差（MAE）を，x 座標，y 座標別々に求めて平均し，精度を評価した。

3　結果と考察

Table 1 はテストデータの全 2332 フレームにおける舌，上唇，下唇，口蓋，上咽頭，下咽頭，喉頭の 7 部位において 2 つのモデルで算出した誤差である。推定した輪郭点と正解点の MAE の平均は，DNN で 1.07 pixel，CNN で 1.28 pixel と，DNN で精度が高かった。

Fig. 3 は誤差が平均的なフレームにおける DNN で推定した輪郭線と正解の輪郭線との比較である。/a/では下唇と喉頭蓋付近などが，/i/では舌の上面と喉頭付近などで誤差が見られたが，おおむねよく一致した。

以上より，推定誤差が約 1 pixel であり，視覚的にも推定した輪郭は正解の輪郭によく一致することから，本研究では音声から調音運動を十分な精度で逆推定できたといえる。

Table 1 各部位の推定精度（単位：pixel）

	DNN	CNN
舌	1.73	1.79
上唇	0.56	0.80
下唇	1.16	1.48
口蓋	0.85	1.27
上咽頭	0.78	1.11
下咽頭	1.01	1.03
喉頭	1.37	1.46
平均	1.07	1.28

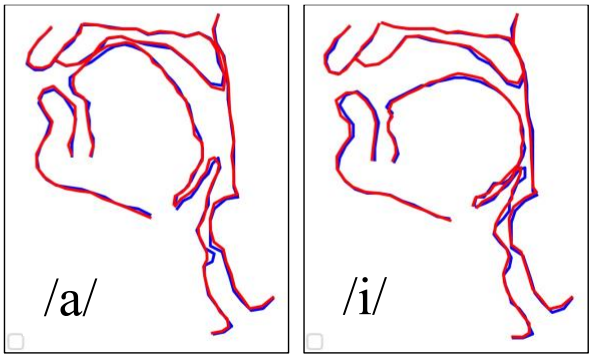

Fig.3 推定した輪郭（青），正解の輪郭（赤）

4　まとめ

本研究では，DNN と CNN の 2 種類のネットワークモデルを構築し，音響特徴量から調音運動の推定を試みた。2 つのモデルの推定誤差を比較した結果，DNN で精度が高く，推定した輪郭点と正解点との平均 MAE は 1.07 pixel であった。また誤差が平均的なフレームで推定した輪郭と正解の輪郭を可視化して比較した結果，おおむねよく一致した。

今後は生成 AI なども用いて精度の向上を図るとともに，現状の単一話者における推定モデルから，任意の話者での推定も可能となるよう多話者への適用についても検討する必要がある。また音素や音素環境の違いが誤差に及ぼす影響についても検討する。

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] Udupa *et al*, Interspeech, 625-629, 2022.
[2] 脇田ら，音講論（秋），927-928, 2023.
[3] S. F. Boll, IEEE Transactions on acoustic, speech, and signal processing, vol. ASSP-27, no.2, pp. 113-120, 1979.
[4] M. Morise *et al*, IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016.
[5] 藤澤ら，音講論（秋），1015-1016, 2022.