音声から調音状態への逆問題における非一意性に関する検討*

☆梶浦一真，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状態を逆推定する音声-調音マッピングの研究が行われており[1]，音声の合成や認識，言語学習や発話治療など様々な分野への応用が期待されている。同一の調音状態に対して音声は一意に生成される。一方，同一とみなすことのできる音声を生成可能な調音状態は多数存在すると考えられている。この問題について，Qin らによる先行研究[2]では英語において，LPC 係数と X 線マイクロビームデータベース（XRDB）を用いて検討を行っている。その結果，逆問題においても多くの場合一意性を示したと報告している。本研究では，Qin らと同様に同一の母音を発声している rtMRI 動画のフレームを音響クラスタリングにより集め，同一のクラスタに属する舌形状の分布が単峰性か多峰性か判別して，逆問題における非一意性について基礎的な検討を行った。

2　材料と方法

2.1 rtMRI 撮像・録音

本研究では，60 代日本人男性 1 名による ATR503 文の発話を ATR-Promotions の Siemens 製 MAGNETOM Prisma fit 3T を用いて rtMRI で動画として撮像し，同時に音声も収録した。フレームレートは 27 fps，解像度は 1×1 mm，スライス厚は 10 mm，画像サイズは 256×256 pixel であった。

2.2 データセット

rtMRI 動画の各フレーム（時間間隔 37 ms）から舌の輪郭を 40 点の点群として抽出した。これは脇田ら[3]と同一のデータで，各点は x, y 座標から成るため，各フレームの舌輪郭は 80 次元のデータである。

rtMRI 撮像と同時に収録した音声から FRCRN 法[4]により MRI 装置のノイズを除去した。その音声から，rtMRI 動画の各フレームの中心時刻における第 1〜3 フォルマント（F1〜F3）と，0 次を除く 1〜24 次のスペクトル包絡（MGC：Mel Generalized Cepstrum）を抽出し，合計 27 次元の音響特徴量とした。なお各フレームにおける中心時刻の MGC は WORLD [5]を用いて 5 ms の間隔で抽出したのち線形補間により求めた。これにより，動画のフレームごとに 80 次元の舌輪郭と 27 次元の音響特徴量のデータセットを構築した。

2.3 クラスタリングの前処理

音声波形に音素と rtMRI 動画のフレーム番号を Praat [6]でラベリングした。そして，音素ラベルを参照して各母音に対応するフレームを集計した。その結果，各母音のフレーム数は，/a/：5546，/i/：3140，/u/：2579，/e/：2931，/o/：5327 となった。

2.4 音響クラスタリング

各フレームの音響特徴量に対して式(1)によって求まる距離 d により，母音ごとに集計された全てのフレームをクラスタリングした。

$$d = \sum((\ln(Fi) - \ln(Fci))^2) + a\sum((Cj - Ccj)^2), \quad (1)$$

ここで Fi（i=1, 2, 3）はフォルマント，Fci は Fi の平均，Cj（j=1,2, …, 24）は MGC，Ccj は Cj の平均，a は重み係数で a = 0.01 である。この距離 d に基づき，python の scikit-learn ライブラリの k-means クラスタリングを用いて各母音で同一の音が集まるよう分類した。その結果，各母音は/a/ : 199，/i/ : 115，/u/ : 162，/e/ : 99，/o/ : 215 のクラスタに分類された。ここで，江口[7]は母音フォルマント弁別域値が中心周波数から 6%の範囲であることを報告している。そのため本研究では，同一の音をクラスタ内における Fci からの誤差が 6%未満と定義し，各クラスタ内で誤差の大きいフレームは削除した。

---

* Investigation of non-uniqueness in the inverse problem from speech sounds to articulatory states, by KAJIURA, Kazuma, WAKITA, Mako, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (National Institute for Japanese Language and Linguistics).

2.5 舌輪郭クラスタリング

まず, 2.4 節で得られた同一の音が集まった各クラスタに対して，scikit-learn ライブラリの Mean-Shift クラスタリングを用いて単峰性を示していることを確認した。次に，クラスタの中でも音響特徴量の平均二乗誤差が比較的小さく，フレーム数が大きいクラスタに着目し，同じく Mean-Shift クラスタリングを用いて音響特徴量に対応する舌輪郭を分類した。これはベクトルの点群から密度のノード（極大点）を探すものであり，Qin ら[2]と同様の手法である。

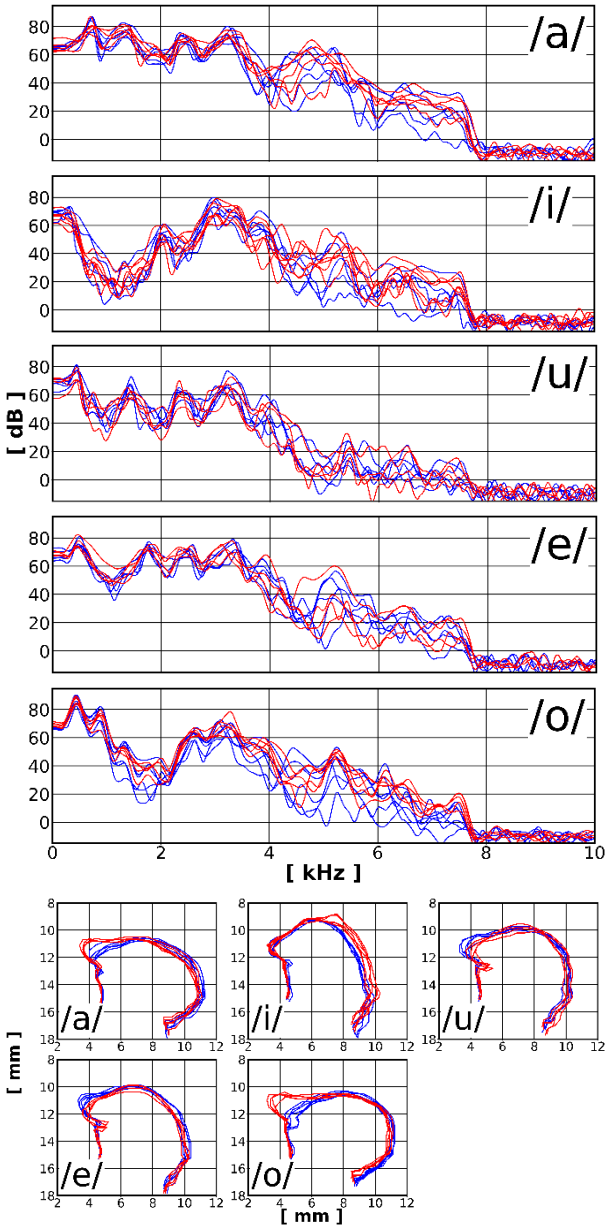

Fig. 1 5 母音のスペクトル（上）と舌形状（下）

3　結果と考察

Fig. 1 は各母音で同一の音として分類されたスペクトルと対応する舌輪郭を，舌輪郭のクラスタに基づいて赤と青に色分けして代表的な 5 フレームをプロットしたものである。これは，単峰性のスペクトルのクラスタが，声道形状では多峰性を示すことを意味しており，同一の音を違ったアプローチで生成している可能性を示唆する。なお，Table1 は Fig. 1 で用いたクラスタに含まれるフレーム数と Fc1～Fc3 の周波数である。

Table 1 Fig. 1 で用いたクラスタ

母音	フレーム数	Fc1	Fc2	Fc3
/a/	110	701.7	1387.4	2405
/i/	81	289.7	2049.7	2916.8
/u/	70	432.8	1431.9	2362
/e/	95	489.9	1745	2424.3
/o/	100	442.5	867.8	2602.5

4　まとめ

本研究では rtMRI 動画から抽出した舌輪郭と音声を用いて，音声から調音状態への逆問題における非一意性の検討を行った。その結果，音響的には同一のクラスタに属するが，舌形状に着目すると複数のクラスタに分類される場合があった。

今後の課題として, DNN を用いたノイズリダクションである FRCRN 法は，学習データに依存して特徴量に特定の値が出やすい可能性が考えられること，仰臥位の rtMRI 撮像では立位と重力の方向が異なるため舌形状が変化している可能性が考えられること，などの影響についても検討する。

謝辞

本研究は JSPS 科研費 24K00071 の助成を受けて実施した。

参考文献

[1] 田口，鏑木，音響学会誌, 77 巻（2），103-111, 2021.
[2] Qin and Carreira-Perpiñán, Interspeech, 74-77, 2007.
[3] 脇田ら，音講論（秋）, 927-928, 2023.
[4] Zaho *et al*, ICASSP, 9281-9285, 2022.
[5] Morise *et al*, IEICE transactions on information and systems, E99-D（7），1877-1884, 2016.
[6] Boersma, Paul, Glot International, 5:9/10, 341-345, 2001.
[7] 江口, Audiology Japan, Vol.15（5），521-522, 1972