音声－調音マッピングの逆問題における非一意性の再検討*

◎梶浦一真，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[1]では，非一意性の問題があると考えられている。しかし，英語音声の LPC 係数と X 線マイクロビームデータベースの声道形状を用いた先行研究[2]では，多くの場合一意性を示した。われわれは，前報[3]で日本語の母音に限定し，音響特徴量と舌形状のクラスタリングを用いてこの逆問題の非一意性を検討したところ，多くの場合に非一意性がみられた。しかし，「同一音声」の定義が限定的であったことから，音声－調音マッピングを考察するには不十分であった。そこで本稿では，同一音声の定義および分析手法を見直し，一部の子音にも対象音素を拡張して再検討を行った。

2　材料と方法

2.1 rtMRI 撮像・録音

60 代日本人男性 1 名による ATR503 文[4]の発話を ATR-Promotions の MAGNETOM Prisma fit 3T を用いてリアルタイム MRI（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 データの前処理

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[5]と同一のデータで，各点は x, y 座標からなるため，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[6]により MRI 装置のノイズを除去した。その音声波形に音素と rtMRI 動画のフレーム番号を Praat [7]を用いてラベリングした。そして，音素ラベルを参照して音素ごとに対応する音声フレームを収集した。なお収集したフレームは，全区間が当該音素に対応しており，フレーム中心時刻の前後 10 ms においてメル一般化ケプストラム（MGC：Mel Generalized Cepstrum）の大きな変化のないものとした。その結果，各音素の音声フレーム数は，/a/ : 3032，/i/ : 1483，/u/ : 796，/e/ : 1464，/o/ : 1988，/m/ : 283，/s/ : 700 であった。これらの収集した音声フレームのスペクトルを予備的に分析し，この話者では Table 1 で示す周波数付近に第 1〜3 ピークが出現すると予測した。そして，全てのフレームで 24 次の LPC を求め，Table 1 で示すピーク周波数に最も近い極の周波数をピークとして抽出した。また，rtMRI フレーム中心時刻の 0 次を除く 1〜24 次の MGC を WORLD[8]を用いて抽出した。

Table 1 予測されたピーク周波数（Hz）

音素	第1ピーク	第2ピーク	第3ピーク
/a/	700	1300	2400
/i/	320	2000	2700
/u/	370	1400	2300
/e/	480	1700	2400
/o/	440	950	2400
/m/	250	2500	-
/s/	1500	5000	-

2.3 同一音声のグルーピング

本研究では，知覚的類似性と音響特徴量空間における単峰性を併せ持つ以下の(1)〜(3)の条件を満たすフレームの音声を同一であると定義した。

(1) ある音素の全てのフレームの音声を一つずつ基準音声とし，それ以外のフレームの音声の各ピーク周波数が Lijzenga [9]による聴覚弁別域（JND：Just Noticeable Difference）未満であること。(2) (1)で収集したフレームの中で，MGC 特徴量空間における基準音声とのユークリッド距離が小さいこと。(3) (2)で収集したフレームの音声を MGC 特徴量空間で mean-shift クラスタリングし，基準音声と同一のクラスタに属すること。なお，(1)における各ピークの単独 JND は，第 1 : 6%，第 2 : 6%，第 3 : 10%とした。

---

* Revisiting the Non-Uniqueness in Acoustic-to-Articulatory Inversion, by KAJIURA, Kazuma, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (National Institute for Japanese Language and Linguistics).

2.4 舌形状クラスタリング

同一音声に分類された全てのフレームの舌形状を主成分分析して 2 次元に圧縮した。その第 1, 2 主成分スコアを mean-shift クラスタリングし, フレーム数が 10 以下のクラスタは除去して，舌形状が単峰性か多峰性かを確認した。以下の Fig.1 は 2.3 節，2.4 節の処理を模式的に示したものである。

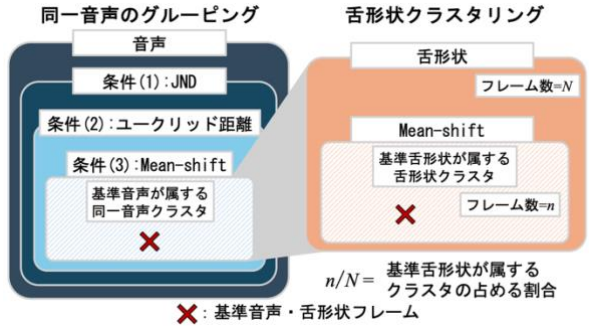

Fig.1 同一音声のグルーピングと舌形状のクラスタリング

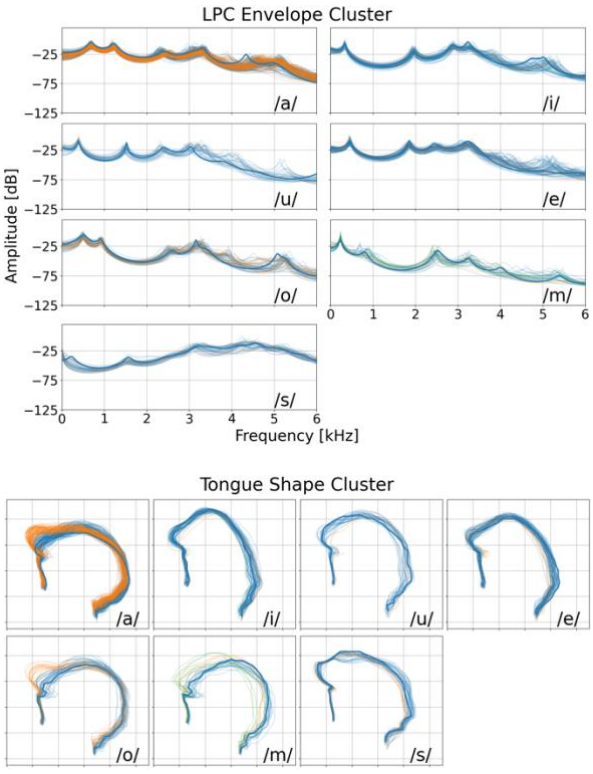

Fig. 2 各音素のスペクトル(上)と舌形状(下)

3 結果と考察

Fig. 2 は各音素の舌形状をクラスタで色分けし，そのフレームの音声スペクトルも同色としたものである。舌形状クラスタリングの結果，舌形状はどの音素でも複数のクラスタに分類される多峰性を示した。

Fig. 1 で示す基準舌形状が属する同一舌形状のクラスタ（フレーム数 n）が，同一音声のクラスタ（フレーム数 N）に占める割合 n/N を，全ての同一音声のクラスタで求め，その平均値を音素ごとにまとめた Table 2 より, /i/, /u/, /e/, /s/では，割合は 95%以上であった。すなわち，これらの音素では，音声と舌形状の一対一対応は 95%以上の割合でみられることを意味する。一方で/a/，/o/，/m/ではこの割合が低く，要因として，/a/，/o/は狭めが舌の後方にあり，/m/は鼻音であるので，舌先端の形状が音響特性に与える影響が小さく，多様な先端形状が許容されることが考えられる。

Table 2 各音素で基準舌形状の属するクラスタが同一音声のクラスタに占める割合

音素	同一音声クラスタの総数	n/Nの平均値［%］
/a/	2415	77.5
/i/	1015	96.7
/u/	174	96.0
/e/	1128	98.6
/o/	1226	72.2
/m/	75	67.5
/s/	460	98.3

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，今回使用したデータでは，どの音素でも非一意性はみられたものの，/i/，/u/，/e/，/s/の音素については高い割合で一対一対応が確認された。一方で/a/，/o/，/m/は非一意性の傾向が強く見られた。

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（2）, 103-111, 2021.

[2] Qin and Carreira-Perpiñán, Porc. Interspeech, 74-77, 2007.

[3] 梶浦ら，音講論（秋） , 947-948, 2024.

[4] 小林ら，音響学会誌，48 (12) , 888-893, 1992.

[5] 脇田ら，音講論（秋）, 927-928, 2023.

[6] Zaho *et al.*, Porc. ICASSP, 9281-9285, 2022.

[7] Boersma, Glot International, 5:9/10, 341-345, 2001.

[8] Morise *et al.*, Porc. IEICE, E99.D (7), 1877-1884, 2016.

[9] Lijzenga, Discrimination of Simplified Vowel Spectra, chapter5-IV, 87-91, 1997.