rtMRI動画から発話器官の輪郭を抽出する学習器の生成に関する検討*

☆後藤翼，竹本浩典（千葉工大），
北村達也（甲南大），能田由紀子，前川喜久雄（国語研）

1　はじめに

われわれは，日本語調音音声学の精緻化のために，18 名の被験者の発話運動をリアルタイム MRI（rtMRI）動画で記録してデータベースを構築してきた[1]。そして，発話運動を定量的に分析するために，動画の各フレームから発話器官の輪郭を点群データとして抽出する手法を開発してきた[2]。これまでのところ，1 名の被験者の動画から 20 フレームほど手動で輪郭を抽出（トレース）して学習器を生成すれば，同じ被験者の 2 万を超えるフレームから精度よく自動的に輪郭を抽出できることが明らかになった[3]。

しかし，予備実験によれば，ある被験者で作成した学習器は，必ずしも別の被験者に適用できなかった。すなわち，発話器官の形状が類似している被験者には適用できたが，そうでなければ適用できなかった。しかし，どの程度類似していれば適用できるのかはっきりしなかった。そこで，本稿では，発話器官の概形をクラスタリングし，これに基づいて教師データを選定し，全ての被験者に適用できる学習器を生成する方法を検討したので報告する。

2　材料と方法

2.1　rtMRI 動画

材料は，日本人成人男性 12 名（M1～M12），女性 6 名（F1～F7，ただし F2 は欠番）によるキャリア文「これが○○型」による 2 モーラ語発話 20 文を含む rtMRI 動画 18 本（1 被験者につき 1 本）とした。なお，装置は（株）ATR-Promotions に設置された Simens 製 MAGNETOM Prisma fit 3 で，空間解像度は 1×1×10mm，フレームレートは 13.8 fps，フレーム数は 512，撮像時間は 37 秒であった。

2.1 機械学習

機械学習には顔認証等に用いられるライブラリ Dlib [4]を用いた。アルゴリズムはランダムフォレスト[4]である。それぞれの被験者の動画の 19 フレームから舌の輪郭を 40 点でトレースして教師データとした。なお，トレースしたフレームは被験者で共通の音韻とした。

2.2 クラスタリング

本稿では，Fig. 1 で示す赤い四角形を発話器官の概形と定義し，全被験者で計測した。四角形の左端は鼻の頂点（点線），右端は第二頸椎の右下の角（1），上端は前鼻棘（2）と後尾棘（3）の中点，下端は第六頸椎の左上の角（4）である。この概形を最小分散法でクラスタリングした。

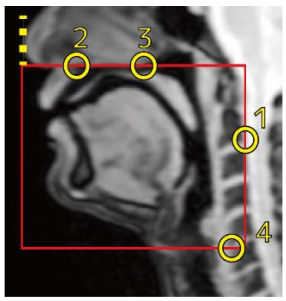

Fig. 1 発話器官の概形

2.3 評価方法

クラスタリングの結果に基づき，18 名の被験者から 8 名を選んで教師データとし，残りの 10 名からそれぞれ 10 フレームを用いて機械学習による輪郭抽出の誤差[3]を計算した。この 18 名から 8 名を抽出する組み合わせと平均誤差・標準偏差を検討した。

3　結果・考察

Fig. 2 は発話器官の概形による被験者のクラスタリングの結果である。18 名の被験者は大きく 3 つのクラスタに分類された。Fig. 2 で示すように，赤が発話器官が小さい女性型（女性 5 名），青が発話器官が大きい男性型（男性 7 名），緑がそれ以外の混合型（男性 4 名，女性 3 名）である。

まず，女性型，男性型，混合型を中心とす

---

* Examination of building learners which extract contours of the speech organs from rtMRI, by GOTO, Tsubasa, TAKEMOTO, Hironori (Chiba Institute of Technology), KITAMURA, Tatsuya (Konan Univ.), NOTA yukiko, and MAEKAWA, Kikuo (NINJAL)

る 8 名を教師データとしたときの機械学習による輪郭抽出の平均誤差・標準偏差を検討した。それぞれの教師データの組み合わせを Table 1 の C1，C2，C3 で示す。Table 1 の各列の最下段 2 行が平均誤差・標準偏差である。この表が示すように，C3 で平均誤差・標準偏差が最も小さかった。これは，混合型が発話器官の概形の多様なバリエーションを含むため，女性型，男性型にも広く適用できるからではないか考えられる。また，この結果は，混合型を中心に微調整を行なえばさらに良い結果を得られることを示唆する。

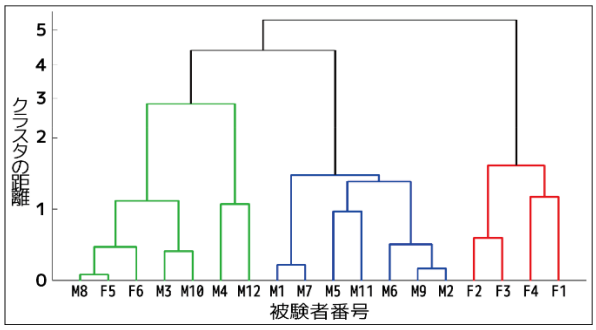

Fig. 2 発話器官の概形のデンドログラム

次に，混合型のクラスタ内で距離が近い 2 名の被験者のうち一方を女性型，男性型のクラスタの被験者と入れ替えることにした。M8 を F1 と入れ替える組み合わせを C4 とし，これに加えて，M10 を M2 と入れ替える組み合わせを C5 とした。その結果，平均誤差が C4 では 0.95，C5 では 0.98 となった。なお，先行研究により，平均誤差が 1.0 以下となると，目視では機械抽出による輪郭抽出は極めて良好といえる[5]。

C3 の M8 を F1 と入れ替えた C4 で平均誤差が減少したのは，F1 が混合型から発話器官の形状が最も遠いためであると考えられる。女性型では四角形の高さ，つまり発話器官の縦の長さが小さい。F1 ではこれが特に顕著であるため，F1 を教師データに含むことで女性型での抽出誤差が小さくなったと考えられる。しかし，C4 の M10 を M2 と入れ替えた C5 では平均誤差は改善しなかった。この理由は C4 に発話器官が大きい男性型の M1 が含まれていたため，M2 を加えても男性型の平均誤差がそれほど小さくならなかったからではないかと考えられる。

もちろん，教師データをこれ以外の組み合わせとすることで，さらに平均誤差が小さくなる場合もありうる。また，学習器生成のパラメーターにおいて，決定木の深さを 2 から 4 に，処理階層の深さを 10 から 20 とすることで C4 の平均誤差が 0.77 と小さくなった。

4 まとめ

18 名の被験者を発話器官の概形を分析したところ，女性型，男性型，混合型の 3 つにクラスタリングされた。これに基づいて教師データとする 8 名の被験者の組み合わせを検討した。その結果，主に混合型と混合型から最も離れている女性被験者を教師データとして学習器を生成したとき，最も輪郭抽出の精度が高くなった。

Table1 教師データとする被験者（○）の組み合わせ（C1～C5）と平均誤差（ME）と標準偏差（SD）。単位は pixel。

	C1	C2	C3	C4	C5
M8			○		
F5			○	○	○
F6			○	○	○
M3			○	○	○
M10			○	○	
M4			○	○	○
M12		○	○	○	○
M1		○	○	○	○
M7		○			
M5		○			
M11	○	○			
M6	○	○			
M9	○	○			
M2	○	○			○
F2	○				
F3	○				
F4	○				
F1	○			○	○
ME	1.17	1.30	1.04	0.95	0.98
SD	1.28	1.52	1.19	0.93	0.92

5 謝辞

本研究は JSPS 科研費 17H02339 の助成を受けた。

6 参考文献

[1] 前川ら，音講論（春），1247-1248，2018.
[2] 後藤ら，音講論（秋），813-814，2018.
[3] 後藤ら，音韻論（春），821-822，2019.
[4] King, Mach. Learn. Res., 1755-1758, 2009.
[5] Takemoto *et al.*, Proc. of Interspeech 2019, 904-908, 2019.