機械学習による複数話者のrtMRI動画における発話器官の輪郭抽出精度の評価*

☆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

近年，発話運動を実時間の動画として記録することが可能になった（rtMRI 動画）。これらの動画から発話器官の輪郭を抽出して発話器官の運動タイミングなどを分析すれば，調音音声学の精緻化が期待できる[1]。

そこでわれわれは，機械学習を用いて半自動的に発話器官の輪郭を抽出する手法について検討を行ってきた[2]。前報では，ある特定の被験者の 1 本の rtMRI 動画のうち 20 フレームから手動で発話器官の輪郭を抽出（トレース）して学習器を生成すれば，同じ被験者の全ての rtMRI 動画の約 2 万フレームからトレースと同程度の精度で発話器官の輪郭を自動的に抽出できることを報告した[3]。

しかし，分析する被験者は男女合わせて 20 名以上である。これらの動画を処理するにあたり，2 つの問題を検討する必要がある。まず，トレース方法の問題である。前報までは，発話器官の輪郭を等間隔の点群でトレースしていた。この方法では，例えば輪郭上に局所的な凹凸（凹点，凸点）がある場合，正確に点を配置することができなかった。そこで，不等間隔でトレースすることでこの問題を解決し，機械学習による輪郭抽出の精度[3]が向上するか検討する。次に，発話器官の形態や運動には個人差があるが，学習させる発話器官の形状の種類と数を共通にできるかという問題である。これを機械学習による輪郭抽出の精度[3]で評価する。本稿では，これら 2 つの問題について，舌形態の個人差が大きな 3 名の被験者で検討したので報告する。

2　方法

2.1 rtMRI 動画

被験者は日本人成人男性 2 名（話者 M1，M2）と女性 1 名（話者 F1）でいずれも標準語話者である。本稿では，これら 3 名の被験者が同一の発話タスクを行った 1 本の rtMRI 動画を分析対象とした。発話タスクは 20 種類の 2 モーラ語のキャリア文「これが○○型」による発話である。これらの動画は，ATR-BAIC に設置された Siemens 製 MAGNETOM Prisma fit 3T で撮像し，解像度 1 mm，スライス厚 10 mm，フレームレート 14 frame/sec で約 37 秒撮像され，512 フレームからなる。なお，以下ではこれらの動画を話者ごとに，動画 M1，M2，F1 と呼称する。

2.2 教師データとするフレームの選定

機械学習の教師データにするために，舌が極端な形状をとるフレームと，軟口蓋，硬口蓋, 咽頭壁と接触するフレームなどの計 19 フレームで舌の輪郭をトレースした。これらのフレームは/a/, /i/, /o/, /g/, /h/, /k/, /n/, /r/, /t/などの音素が対応していた。なお，これらのフレームは 3 名の被験者で共通とした。

2.3 等間隔・不等間隔の点群によるトレースと機械学習による輪郭抽出

まず，3 名の被験者の発話器官の輪郭を前節で述べた 19 フレームで 40 点の等間隔の点群としてトレースした（等間隔トレース）。次に，40 点の等間隔トレースを行い，点を配置できなかった凹点，凸点などに一部の点を移動して微調整した（不等間隔トレース）。そして，2 種類のトレースに対して機械学習ライブラリ Dlib [4]により別々に学習器を生成し，動画 M1，M2，F1 の全てのフレームで輪郭抽出を行った。なお，機械学習のアルゴリズムはランダムフォレスト[4]である。

2.4 輪郭抽出精度の評価

教師データに含まれない，3 名の被験者の動画で共通した音素に対応する 10 フレームで等間隔・不等間隔トレースを行い，輪郭抽出精度の評価の基準とした。これらのフレー

* Examination of edge detection of speech organs from real-time MRI movie for multiple speakers by a machine learning method, by GOTO, Tsubasa, TAKEMOTO, Hironori (Chiba Institute of Technology), KITAMURA, Tatsuya (Konan Univ.), NOTA yukiko, and MAEKAWA, Kikuo

ムで機械学習によって得られた輪郭上の点と，トレースによって得られた輪郭で最も近傍な線分との最短距離の平均値をピクセル単位で求めて誤差とした（図 1）。この誤差による評価を 3 名の被験者における等間隔・不等間隔トレースの両方で行った。

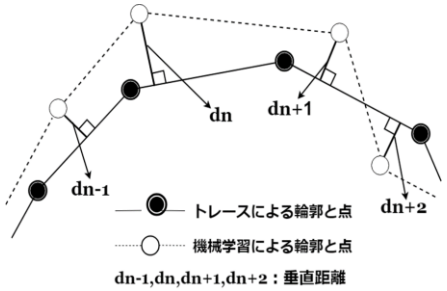

図 1. トレースした輪郭線に対する機械学習による輪郭抽出点の誤差の算出方法

3　結果と考察

まず，等間隔・不等間隔トレースに基づく機械学習による輪郭抽出（以下，等間隔・不等間隔トレースによる抽出）の精度を比較する。表 1 で示すように，不等間隔トレースによる抽出の平均誤差と標準偏差は，3 名の被験者すべてで等間隔トレースによる抽出より小さくなった。すなわち，等間隔トレースによる抽出より，不等間隔トレースによる抽出の方が精度が高いことが示された。

図 2 は動画 M2 の輪郭抽出精度の評価に用いた 233 フレーム（/a/に対応）におけるトレースと機械学習で抽出した輪郭を表す点（輪郭点）である。誤差が示すように，不等間隔トレースの方が精度が高い。これは，等間隔トレースでは凸点や凹点を正確にトレースできないこと（図 2 左の黄色の円）が輪郭の学習に悪影響を与えたためと考えられる。逆に，不等間隔トレースによる抽出では，これらの凹点，凸点は正確に抽出された（図 2 右の空色の円）。これは，不等間隔トレースにより凹点，凸点が正確にトレースされたことが輪郭の学習に反映されたためと考えられる。

最後に，被験者間の輪郭抽出精度のばらつきについて検討する。表 1 で示すように，等間隔・不等間隔トレースによる抽出のいずれも，被験者間のばらつきは小さい。これは，2.2 節で述べた教師データとするフレームの選択が複数の被験者にも有効であったことを意味する。また，不等間隔トレースによる抽出は等間隔トレースによる抽出より，さらに被験者間のばらつきが小さいという結果が得られた。この原因ははっきりしないが，不等間隔トレースでは，被験者の発話器官の形態の個人差によらず正確にトレースできることが要因の一つではないかと考えられる。

表 1 動画 M1, M2, F1 における機械学習による輪郭抽出の平均誤差（標準偏差）[pixel]

	等間隔トレース	不等間隔トレース
M1	0.76 (0.67)	0.68(0.66)
M2	0.81(0.77)	0.70(0.67)
F1	0.75(0.72)	0.72(0.66)
平均	0.77(0.75)	0.70(0.70)

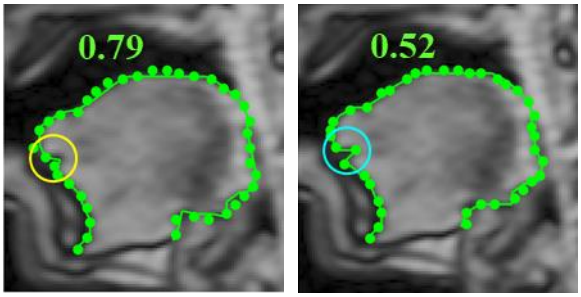

図 2. 動画 M2 の第 233 フレームにおける等間隔（左）・不等間隔（右）トレースの輪郭線と抽出した輪郭点。数値は平均誤差 [pixel]。

4　おわりに

本稿では，不等間隔の点群で発話器官をトレースすることにより，機械学習による輪郭抽出の精度が向上することを示した。また，発話器官の形態に個人差があるが，学習器に学習させる形状の種類と数を 3 人の被験者間で共通にしても，輪郭抽出の精度に差はないことを示した。

謝辞

本研究は JSPS 科研費 17H02339 の助成を受けた。

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., 10, 1755-1758, 2009.