機械学習によるrtMRI動画における発話器官の輪郭抽出精度の評価*

☆後藤翼，萩原裕也，濱中彩夏，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

われわれは日本語の調音音声学を精緻するために，多数の被験者の多様な発話運動をリアルタイム MRI 動画（以下 rtMRI 動画）で記録し，発話器官各部の運動タイミングなどを定量的に分析しようとしている[1]。そのためには発話器官の輪郭を点群データに変換（輪郭抽出）する必要がある。しかし，われわれは 1 被験者につき 2 万を超えるフレームを分析しようとしているため，手動による輪郭抽出（トレース）は非現実的である。

そこで，われわれは機械学習による輪郭抽出に着手した[2]。これは，特定のフレームと，そのトレースした輪郭の対を教師データとして学習させ，得られた学習器により全てのフレームから輪郭を同じ数の点群として抽出するものである。その結果，目視では妥当な輪郭が抽出されることを確認した[2]。また，抽出精度を評価するために，学習に用いていないフレームのトレースで得られた輪郭と，機械学習で得られた輪郭の対応点の平均距離を誤差として求めた。しかし，この方法では，機械学習で得られた点が輪郭上にあっても，対応点から離れていれば誤差が大きくなるため，評価法として不十分であった。そこで本稿では，この点を改良した評価法を提案し，様々な評価を行ったので報告する。

2　輪郭抽出方法

2.1 rtMRI 動画

被験者は日本人成人男性 1 名（標準語話者），MRI 装置は ATR-BAIC に設置された Siemens 製 MAGNETOM Prisma fit 3T である。rtMRI 動画は 55 本（動画 4～58）あり，それぞれ約 20 種類の 2 モーラ語をキャリア文「これは○○型」により発話した際に撮像したものである。いずれも解像度は 1 mm，スライス厚は 10 mm で 14 frame/sec の速度で撮像され，512 フレームで構成されている。このうち，動画 15 の特定のフレームを学習に使用し，動画 4，10，15，20，30，45 で機械学習による輪郭抽出と評価を行った。

2.2 機械学習による輪郭抽出

輪郭抽出は，舌，口唇・下顎，軟・硬口蓋，咽頭後壁・披裂部，喉頭蓋・声帯の 5 つの部位で行った。動画 15 の特定のフレームにおけるこれらの部位を，発話器官の解剖学的知見を有し，トレース経験が豊富な 3 人の熟練者がトレースして教師データを作成した。Table 1 に示すように，輪郭を構成する点数，トレースを行うフレーム数は部位ごとに異なる。点数は輪郭線の長さと形状の複雑さに応じて設定し，フレーム数は他の部位と接触するパタンなど（例：舌と歯茎，硬口蓋，軟口蓋の接触）に応じて設定した。後者に関しては，予備実験により，2 つの部位が接触するフレームで機械学習による輪郭抽出が不良であり，これらのフレームを重点的に学習させる必要があることが判明したためである。

Table 1 Number of points and frames for each edge

部位	点数	フレーム数
舌	40	19
口唇・下顎	40	14
軟・硬口蓋	40	17
咽頭後壁・披裂部	30	10
喉頭蓋・声帯	30	11

このようにしてトレースした輪郭を，機械学習ライブラリ Dlib [3]に教師データとして与えて輪郭抽出を学習させた。得られた学習器を用いて，全ての動画で輪郭抽出を行った。

2.3 輪郭抽出精度の評価

輪郭抽出精度を評価するための誤差の計算

* Accuracy evaluation of edge detection of speech organs from real-time MRI movie by a machine learning method, by GOTO, Tsubasa, HAGIHARA, Yuya, HAMANAKA, Sayaka, TAKEMOTO, Hironori (Chiba Institute of Technology), KITAMURA, Tatsuya (Konan University), NOTA Yukiko, and MAEKAWA, Kikuo (NINJAL)

は，以下のように行った。まず，あるフレームで 5 つの部位それぞれに対して，機械学習によって得られた輪郭の点と，熟練者の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（Fig. 1 参照）の平均値をピクセル単位で求めて各部位の誤差とした。次に，5 つの部位の誤差を平均してそのフレームの誤差とした。さらに，動画 15 以外のすべての動画の第 1，256，512 フレームの誤差を平均して，その動画の誤差とした。この誤差は，機械学習によって得られた点が輪郭上にずれている場合が多ければ小さくなり，輪郭から垂直方向にずれている場合が多ければ大きくなる。なお，動画 15 では，第 1，512 フレームを学習に用いたので，それ以外の学習に用いていない 24 フレームで詳細に評価した。

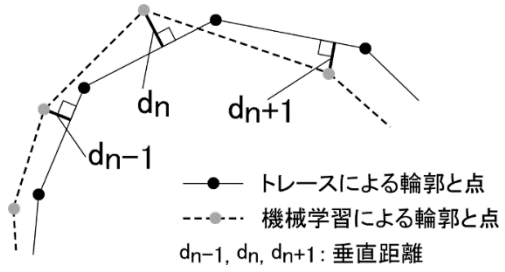

Fig. 1 Vertical distances from the points detected by a machine learning to the edge line extracted by manual tracing.

3　結果と考察

まず，動画 15 での精度評価の結果を述べる。熟練者のトレースによる輪郭抽出と機械学習による輪郭抽出の誤差は 0.91 であった。この値を評価するために，3 名の熟練者が抽出した輪郭の差を誤差として計算したところ，平均 1.15 であった。すなわち，ある熟練者の輪郭抽出を学習した学習器は，別の熟練者より本人に近い輪郭を抽出できることを示している。なお，熟練者と非熟練者（アルバイトを想定し，簡単なインストラクションを受けた 16 名の学生）の誤差は平均 1.28 であった。

次に，動画 15 で作成した学習器を他の動画の輪郭抽出に適用した結果について述べる。各動画の撮像の間には，装置の調整などによる数秒から数十秒にわたる不等間隔のインターバルがあるため，その間の不随意な体動によって頭部位置が変化し，断面形状が変化することがある。そのため，動画 15 で作成した学習器は，必ずしも他の動画には適用できない可能性がある。特に，動画 15 から時間的に隔たって撮像された動画ほど誤差が大きくなりうる。Table 2 は動画 15 との時差と動画の誤差の表である。最大で 30 分程度の時差があるが，どの動画でも誤差は動画 15 と同程度であった。つまり，特定の動画で生成した学習器は，その被験者による全ての動画に有効である可能性が示唆された。なお，Fig. 2 は動画 30 の第 256 フレームで，左はトレースによる輪郭，右は機械学習による輪郭であり，このフレームの誤差は 1.05 ピクセルであった。

Table 2. Estimation errors (mean vertical distances) and time lags for movies

movie	4	10	15	20	30	45
time lag	-9:51	-4:23	0:00	4:30	14:05	30:37
error	0.97	0.90	0.91	0.94	1.11	0.88

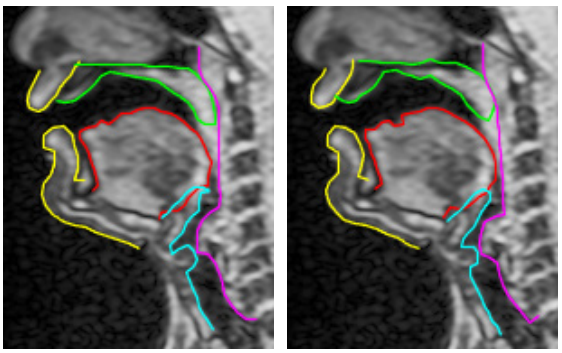

Fig. 2 Edges of speech organs on the 256th frame of movie 30 extracted by a manual trace (left) and a machine learning (right).

4　おわりに

本稿では，機械学習による rtMRI 動画における発話器官の輪郭抽出精度の評価法を提案し，評価を行った。その結果，最大で 20 フレームのトレースを学習させることで，2 万以上のフレームから熟練者と同程度の精度で輪郭を抽出できることが明らかになった。

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] King, Mach. Learn. Res., 10, 1755-1758, 2009.