機械学習によるrtMRI動画における発話器官の輪郭抽出方法の検討*

☆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），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の精緻化のために，発話運動のリアルタイム MRI 動画（以下 rtMRI）の収録を進めている[1]。各被験者は単独モーラ，主要 2 モーラの組み合わせなど約 900 発話を行い，その正中矢状断面における発話運動を動画として記録する。これまでに標準語話者の男性 6 名，女性 2 名，近畿方言話者の女性 2 名で撮像を行ってきた。

この動画を適切に処理すれば，発話器官各部の運動タイミングなどを定量的に分析できる。しかし，rtMRI の動画は組織の輪郭が不鮮明であるため，定量的な分析を行うためにはフレームごとに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。これにより，例えば舌と硬口蓋の接触を判定できる。ところが，フレーム数が膨大であるため，手動による輪郭抽出は現実的でない。

Naranayan らの rtMRI を用いたデータベース[2]では，周波数領域で輪郭抽出する方法が検討されている[3]。しかし，その手法は多大な計算時間が必要である[3]。そこで，われわれは機械学習による発話器官の輪郭抽出を試みた。本稿では，処理の概要，現在までに得られている結果，および誤差について検討したので報告する。

2　輪郭抽出方法

2.1 rtMRI データ

対象となるデータは，日本人成人男性 1 名（標準語話者）がキャリア文「これが～型」により 2 モーラ語（ク＋ハ行，サ行，マ行，シャ，シュ，ショおよびケカ，ケキ）20 文を発話した動画である。動画は 14 frame/sec の速度で約 37 秒間撮像され，合計 512 フレームで構成されている。解像度は 1 pixel が 1×1 mm，スライス厚は 10 mm である。なお，MRI 装置は ATR-BAIC に設置された Siemens 製 MAGNETOM Prisma fit 3T を用いた。

2.2 輪郭抽出を行う部位

輪郭抽出は Fig. 1 で示す以下の 5 つの部位で行った。すなわち，舌（赤），口唇・下顎（青），軟・硬口蓋（緑），咽頭後壁・披裂部（マゼンタ），喉頭蓋・声帯（シアン）である。

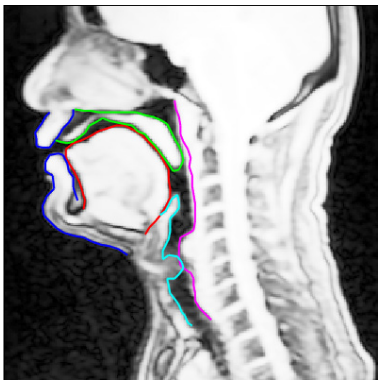

Fig. 1 Five parts of edge detection (See text).

2.3 処理の概要と誤差の検討

処理の概要は以下の通りである。(1) rtMRI のフレームから目視による輪郭抽出（以下トレース）を行って学習データを作成する。(2) これを教師データとして機械学習ライブラリ Dlib [4]に輪郭抽出を学習させる。(3) 得られた学習器を用いて全フレームに対して輪郭抽出を実行する。

(1) について補足する。トレースは MATLAB を用いて作成した GUI プログラムで実施した。例えば，舌の輪郭は，顎舌骨筋の下顎骨への付着部位から下顎骨後面，舌尖，舌背，喉頭蓋上面，顎舌骨筋の舌骨への付着部位に至る輪郭を 30 点でトレースした。また，トレースを行ったのは計 15 フレームで，その内訳は舌先の形状が特殊なフレームを 3，喉頭蓋と舌背が接触しているフレームを 1，舌が歯茎・硬口蓋・軟口蓋のどれかに接触しているフレームを各 3，第 1 フレームと最終

* Examination 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), and MAEKAWA, Kikuo (NINJAL)

フレームであった。その他の部位も目視で同定が容易な点を始点・終点とし，形状の変動を考慮してトレースするフレームと枚数を独立に決定した。その内訳は，口唇・下顎を 40 点（上唇 15 点，下唇・下顎 25 点）で 11 フレーム，軟・硬口蓋を 40 点で 17 フレーム，咽頭後壁・披裂部を 30 点で 6 フレーム，喉頭蓋・声帯を 30 点で 11 フレームであった。

トレースと機械学習による輪郭抽出の誤差の検討は以下のように行った。まず，学習データと別に複数のフレームでトレースを行なった。そして，その各点の座標に対する機械学習による輪郭抽出で得られた対応する点の座標のピクセル距離を誤差として求め，平均値と標準偏差を求めた（Fig. 2）。なお，誤差の検討に用いたフレーム数は，舌が 5，口唇が 3，軟・硬口蓋が 5，咽頭後壁・披裂部が 6，喉頭蓋・声帯が 5 であった。

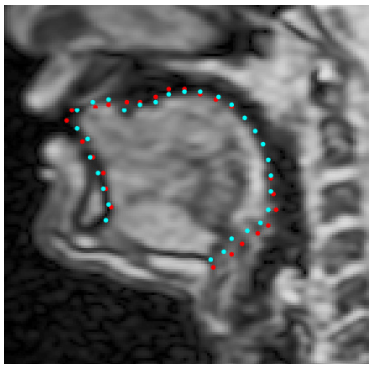

Fig. 2 Cyan points: manually traced, red points: detected by a machine learning.

3　結果と考察

Fig. 3 は動画のあるフレームにおける機械学習によって抽出された発話器官の輪郭である。目視では，このフレームにおける抽出結果はほぼ妥当であるといえる。

Table 1 は部位ごとにトレースと機械学習による輪郭抽出の誤差を示したもので，単位はピクセルである。最も誤差が大きかったのは喉頭蓋・声帯であった。この部位では画質が悪く，輪郭が不明瞭でトレースが困難であった。また，この表には記載していないが，どの部位においても接触(例えば舌と咽頭壁)が生じるフレームで誤差が大きかった。これらのフレームでは，接触により輪郭が不明瞭となりトレースが困難であった。すなわち，トレースが難しい部位で誤差が大きくなる傾向があった。

Table 1 Average error and standard deviation [pixel] for each of speech organs.

部位	平均誤差	標準偏差
舌	2.4	1.8
口唇・下顎	3.6	2.6
軟・硬口蓋	3.3	2.1
咽頭後壁・披裂部	3.9	2.5
喉頭蓋・声帯	4.2	2.5

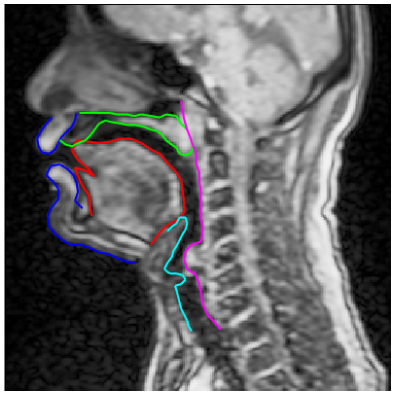

Fig. 3 Edges of speech organs on a frame of rtMRI detected by a machine learning.

4　おわりに

機械学習ライブラリ Dlib によって，rtMRI 動画における発話器官の輪郭抽出を試みた。その結果，目視では良好な結果が得られた。トレースとの誤差は，トレースが困難な部位で大きくなる傾向があった。おそらく，このような部位では，トレースごと，トレースするオペレータごとに輪郭抽出のばらつきが大きいと思われる。機械学習で許容する誤差を検討する上で，そのばらつきの検討も必要であると考えられる。

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] Naranayan *et al.*, JASA 136, 1307-1311, 2014.

[3] Bresch and Naranayan, IEEE Trans. Med. Imaging 28, 323–338, 2009.

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