Basic study of average vocal tract shape based on contours extracted from real-time MRI video

リアルタイム MRI 動画から抽出した声道の輪郭に基づく平均声道の基礎的検討

竹本 浩典（千葉工大）†
天野 沢海（千葉工大）

Basic study of average vocal tract shape based on contours extracted from real-time MRI video

Hironori Takemoto (Chiba Institute of Technology)
Takumi Amano (Chiba Institute of Technology)

要旨

リアルタイム MRI 動画における声道の輪郭は不鮮明であり，声道形状や調音運動を定量的に分析するためには輪郭を点群として抽出する必要がある。しかし，多量の動画フレームから輪郭を手動で抽出することは困難である。そこで，20 名の話者の声道各部の概形を分析し，トレースする話者とフレームを決定して，輪郭を自動抽出する学習器を生成した。そして，この学習器を用いて，全ての話者の動画から輪郭を抽出して抽出精度を評価した。その結果，学習器に含まれていない話者であっても，学習器に含まれている話者と同程度の精度で輪郭を抽出可能であることが明らかになった。さらに，男女 7 名ずつの安静呼吸時の声道形状の輪郭から平均声道を求め，主成分分析した。その結果，第 1 主成分は声道の大きさ，第 2 主成分は首の角度，第 3 主成分は下顎の前後方向の大きさと口の開きを表すことなどが明らかになった。

１．はじめに

日本語の調音音声学を精緻化するためには，多数の日本人話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，定量分析する必要がある（前川他 2018，前川他 2020）。rtMRI 動画では調音器官の輪郭は曖昧であるため，調音運動の定量的な分析を行うためには動画の各フレームから調音器官の輪郭を点群データとして抽出する必要がある。しかし，データベースに含まれる動画は多量であるため，手動による輪郭の抽出（トレース）は現実的ではない。

そこで，2018 年度に機械学習を導入することにより，1 名の話者の 55 本の動画（28,160 フレーム）から人間と同程度の精度で調音器官の輪郭を抽出する手法を確立した（後藤他 2019，Takemoto et al. 2019）。そして 2019 年度には，18 名の話者の舌の概形をクラスタ分析し，これに基づいて選出した 8 名から学習器を生成すれば，全ての話者の動画から舌の輪郭を人間と同程度の精度で抽出できることが明らかになった（後藤他 2020a）。さらに 2020 年度には，定量分析の試みとして，男性話者 12 名，女性話者 6 名の安静呼吸時の声道形状の輪郭データを用いて平均声道を求め，主成分分析やクラスタ分析を行った（後藤他 2020b）。しかし，男性話者数が女性話者数の 2 倍でデータに偏りがあることなどの問題があった。

本研究では，後藤他 2020a の研究手法を調音器官の 5 つの部位に適用し，20 名の話者の rtMRI 動画から輪郭を抽出して精度を検討した。そして，定量分析の試みとして，輪郭を抽

† hironori.takemoto@p.chibakoudai.jp

出した 20 名から男女 7 名ずつ合計 14 名を選出し，安静呼吸時における平均声道を求め，主成分分析などを行うことにより声道形状を検討したので報告する。

２．材料と方法

２.１話者と rtMRI 動画撮像

実験参加者は日本語母語話者 20 名で，男性 13 名（M1~M13），女性 7 名（F1~F7）である。各話者はキャリア文「これが○○型」による 2 モーラ語の発話を行い，約 20 発話ごとに 1 本の rtMRI 動画に記録した。rtMRI 動画は(株)ATR-Promotions に設置された Simens 製 MAGNETOM Prisma fit 3 で撮像した。各動画の空間解像度は 1×1×10 mm，フレームレートは 13.8 fps，フレーム数は 512 であった。

２.２輪郭を抽出する動画と部位

各話者の 34 本の動画から，同一の語群を含む動画を 1 本選択して調音器官の輪郭を抽出した。図 1 は輪郭を抽出する 5 つの部位，すなわち，舌（赤），口唇・下顎（黄），軟・硬口蓋（緑），咽頭後壁（マゼンタ），喉頭蓋・声帯（シアン）である。それぞれ，40 点，40 点（上唇 15 点，下唇・下顎 25 点），40 点，28 点，30 点で構成されている。表 1 は，各部位の始点・終点の解剖学的な位置を示す。

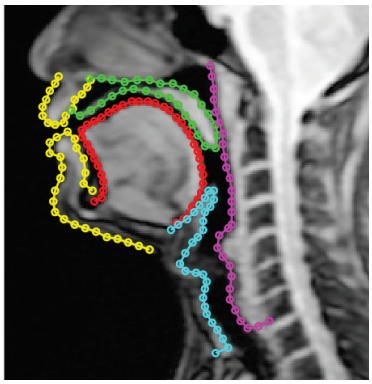

図 1 輪郭を抽出する 5 つの部位

表 1 各部位の始点・終点の位置

部位	始点	終点
舌	二腹筋窩	舌骨の顎舌骨筋付着部
上唇	鼻下点	前鼻棘
下唇・下顎	オトガイ棘前方の海綿骨	上甲状切痕
軟・硬口蓋	前鼻棘	上顎切歯間
咽頭後壁	咽頭後壁上端	第 7 頸椎下端
喉頭蓋・声帯	舌骨の顎舌骨筋付着部	第 7 頸椎下端の 延長方向の胸骨舌骨筋

２.３学習器の生成

まず，20 名の話者から，M13，F7 の 2 名を除いた 18 名で部位ごとに輪郭を抽出する矩形領域と輪郭の終点・始点を決定した。2 名を除外したのは，以下で述べるクラスタ分析の効果を検討するためである。次に，矩形領域の縦・横の大きさと，矩形領域左上を原点とした輪郭の終点・始点の座標をクラスタ分析し，枝の高さなどに基づいて話者を 8 名選出した（図 2)。これは，学習器を生成する上で，偏った形状を過学習させることを避け，でき

るだけ様々な形状を学習させるためである。次に，選出した話者の動画から，表 2 に示すフレームでトレースした。これらのフレームは，各部位の形状のバリエーションと他の部位との接触パタンによって決定した。ここで，/k/(u)は後続母音が/u/である/k/のフレーム，/k/は後続母音が任意であることを示す。また，第 512 フレームは最終フレームである。すなわち，全ての部位で先頭と最終のフレームをトレースした。最後に，トレースしたフレームと点群の座標から学習器を生成し，動画の全フレームから輪郭を自動抽出した。なお，学習器の生成および輪郭の抽出には機械学習ライブラリ Dlib（King 2009）を使用した。

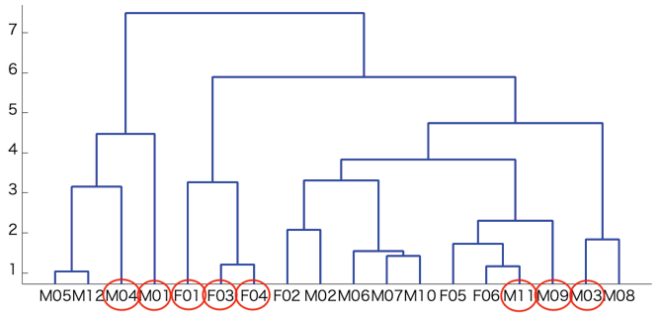

図 2 咽頭後壁の輪郭を抽出する矩形領域のクラスタ分析の結果（赤丸：選出した話者）

表 2 部位ごとにトレースするフレームとフレーム数

部位	1 フレーム	2 フレーム	3 フレーム	合計
舌	/a/, /e/, /k/, /t/, /r/, /n/, 無発話, 第 1・512 フレーム	/i/, /o/, /s/		15
口唇・下顎	/a/, /u/, /e/, /m/, 無発話, 第 1・512 フレーム	/i/, /p/, /o/		13
軟・硬口蓋	/i/, /k/(u), /k/(e), /t/, /m/, 無発話, 第 1・512 フレーム	/a/		10
咽頭後壁	/g/, /m/, 第 1・512 フレーム	/k/, /a/, 無発話		10
喉頭蓋・声帯	/i/, /e/, /o/, 第 1・512 フレーム	/k/, /a/, 無発話	/u/	14

2.4 精度評価

各部位で学習器に含めなかった 12 名の動画を用いて，輪郭抽出の精度を定量的・定性的に評価した。定量評価は先行研究（Takemoto et al. 2019）に基づき，トレースした輪郭を真値として機械学習によって抽出した輪郭の誤差をピクセル値で計算した。なお，この評価は第 256 フレームのみで行った。定性評価は，その部位の輪郭をトレースしたオペレーターが，動画の全てのフレームに対して輪郭抽出の精度を目視により 5 を最高として 5 段階で評価した。

2.5 声道形状の分析

定量分析の試みとして，安静呼吸時の声道形状の性差や個人差を検討した。動画の第 1 フレームは発話直前の安静呼吸時の声道形状である。後藤他 2020b では男性話者 12 名，女性話者 6 名の声道形状を分析したが，全話者の平均声道は男性話者の声道に偏っていると考えられた。そこで，20 名の話者から男女 7 名ずつ合計 14 名の第 1 フレームの輪郭データを用いて声道形状の分析を行った。20 名のうち，女性話者は 7 名（F1～F7）であるので，男女同数とするために，13 名の男性話者からランダムに 7 名（M1, M2, M5, M6, M8, M10, M13）を選出した。

2.2 節で述べたように，全ての話者で抽出した輪郭の始点・終点は解剖学的に対応したラ

ンドマークであり，輪郭線を構成する点の数は等しい。しかし，座標系は話者によって異なり，輪郭線を構成する点群は必ずしも等間隔に配置されていないため，平均声道の生成などの輪郭線の分析処理はできない。

そこで，分析に先立って以下の前処理を行った。まず，全話者の前鼻棘が原点となるように平行移動した。次に，硬口蓋の上縁が水平になるように回転移動した。そして，各輪郭を構成する点を輪郭線に沿って移動させることにより，等間隔の点群（セミランドマーク）となるように変換した。なお，この操作において，ランドマークを増やすために，軟・硬口蓋を構成する 30 点の 15 番目の点が軟口蓋の後端に，咽頭後壁を構成する 28 点の 14 番目の点が披裂部上端に，喉頭蓋・声帯を構成する 30 点の 8 番目の点が喉頭蓋の先端，20 番目の点が声帯の先端となるように調整した。その結果，これらの点と始点・終点を合わせて合計 16 点が全話者で解剖学的に対応するランドマークとなった。図 3 はこれらの前処理を行った M1 と F1 の輪郭線とそれを構成する点群，およびランドマークを示す。

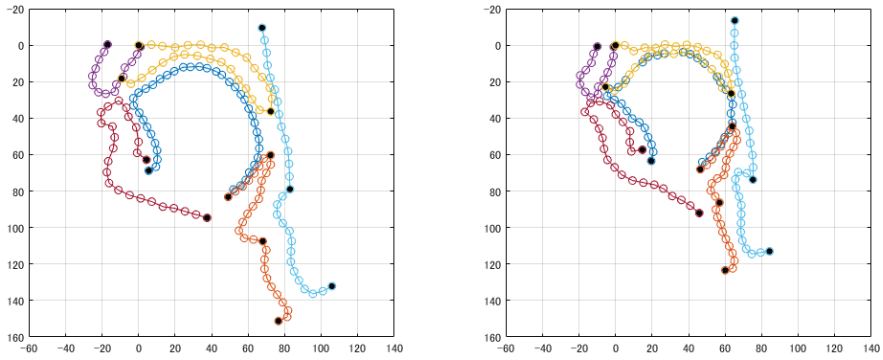

図 3 声道形状分析の前処理（左：M1 の輪郭，右：F1 の輪郭）。黒点はランドマーク。

前処理した 14 名の輪郭線の対応する点ごとに幾何重心を求めることにより，全話者，男性話者，女性話者それぞれの平均声道を求めた。そして，薄板スプライン（e.g., Swiderski 1993）を用いて 3 つの平均声道の形態差を可視化した。また，最小分散法（ward 法）によるクラスタ分析と主成分分析を行った。

３．結果と考察

3.1 輪郭抽出

表 3 は輪郭抽出の精度を定量評価した結果である。先行研究により，誤差が 1.0 pixel 以内であれば，高い精度で輪郭が抽出されたと言える（Takemoto et al. 2019）。誤差が 1.0 pixel 以上の部位を太字で示す。誤差が大きかったのは，喉頭蓋・声帯と咽頭後壁であった。これらの部位は，スライス厚に対して組織が薄いため輪郭が不鮮明になりやすく，これが誤差の大きい要因と考えられた（Takemoto et al. 2019）。また，クラスタ分析に含めていなかった M13 と F7 で特に誤差が大きいという傾向はみられなかった。

表 4 は輪郭抽出の精度を目視により評価した結果で，太字は評価値が 2 以下であることを示す。基本的に表 3 で誤差が小さかった動画は表 4 で評価が高かった。一方，表 3 で誤差が大きかった動画は，必ずしも表 4 で評価が低いとは限らなかった。例えば，表 3 で M7 の喉頭蓋・声帯の誤差は 2.02 で最大であったが，表 N での評価値は 5 であった。これは，定量評価を行った第 256 フレームでは誤差が大きかったが，その他のフレームでは総じて誤差が小さかったためと思われる。

表 3 輪郭抽出の誤差（単位：pixel）

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	F1	F2	F3	F4	F5	F6	F7
口唇・下顎	0.96	0.75	0.56	0.69	**1.23**	-	-	**1.40**	-	0.71	**1.15**	-	0.50	-	-	0.72	**1.02**	-	-	0.99
舌	0.96	0.90	-	-	-	0.78	**1.78**	**1.29**	0.68	0.71	-	-	0.55	-	**1.13**	0.67	-	-	**1.13**	**1.12**
軟・硬口蓋	0.68	**1.51**	-	-	-	0.80	-	0.67	0.94	0.60	0.60	-	0.75	-	**1.10**	**1.15**	0.49	-	-	0.70
咽頭後壁	-	**1.12**	-	-	**1.21**	**1.18**	**1.45**	0.96	-	0.66	-	**1.56**	**1.19**	-	**1.07**	-	-	0.87	0.87	0.70
喉頭蓋・声帯	**1.22**	-	-	0.95	**1.42**	0.88	**2.02**	**1.93**	**1.24**	**1.50**	-	-	**1.03**	**1.27**	-	-	-	**1.75**	-	0.53

表 4 輪郭抽出の精度の 5 段階評価

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	F1	F2	F3	F4	F5	F6	F7
口唇・下顎	4	4	4	4	3	-	-	3	-	5	3	-	4	-	-	4	4	-	-	4
舌	4	4	-	-	-	4	5	4	3	4	-	-	4	-	4	3	-	-	4	3
軟・硬口蓋	3	**2**	-	-	-	5	-	5	3	5	4	-	3	-	3	4	5	-	-	**2**
咽頭後壁	-	4	-	-	4	5	4	4	-	4	-	5	**2**	-	5	-	-	4	5	**2**
喉頭蓋・声帯	5	-	-	5	**2**	4	5	**2**	5	5	-	-	4	**1**	-	-	-	**1**	-	3

なお，この手法では，部位ごとに学習器を生成し，独立して輪郭を抽出するため，隣接する部位で輪郭がオーバーラップすることがある。動画を分析した結果，隣接する二つの部位で動きの大きな方がオーバーラップすることが明らかになった。そこで，オーバーラップを検知し，動きの大きな部位の輪郭を動きの小さな輪郭に自動的に合わせる処理を組み込んだ。これにより，より自然な輪郭を抽出できるようになった（図 4）。

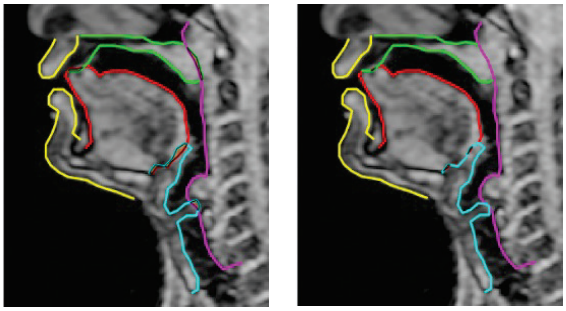

図 4 オーバーラップ解消処理（左：処理前，右：処理後）

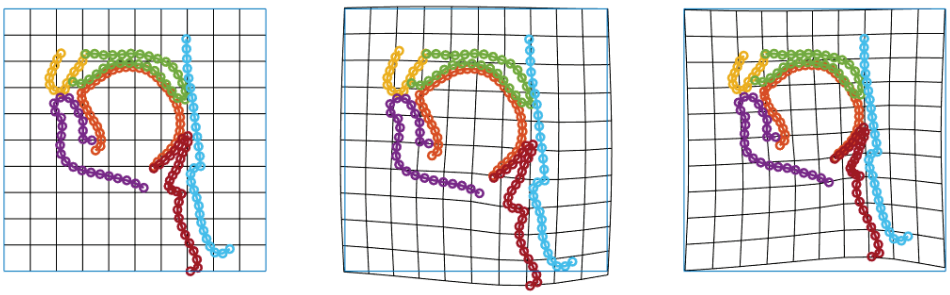

図 5 全話者の平均声道（左），男性話者の平均声道（中），女性話者の平均声道（右）

3.2 声道形状の分析

図 5 は全話者の平均声道で設定した直交格子が，薄板スプライン関数により男性話者の平均声道と女性話者の平均声道にどのように写像されるかを示している。全話者の平均声道に比べて男性話者の平均声道では，全体的に声道形状が拡大し，それは特に下方向に顕著であった。これは，下顎や舌の前下方，声帯や披裂部の後下方への拡大が要因であると考えられる。一方，女性話者ではこれと逆になり，声道形状は全体的に縮小し，それは特に上方向に顕著であった。これは，下顎や舌の後上方，声帯や披裂部の前上方への縮小が要因であ

ると考えられる。これらの差は，男女間の調音器官の大きさ，特に下顎の大きさや喉頭の高さの差を反映していると考えられ，Fitch and Giedd (1999)の結果と一致する。

図 6 は全話者の輪郭形状をクラスタ分析した結果である。まず，大きく男性型と女性型に分かれた。これは，図 5 から声道の大きさが要因であると推測される。さらに，男性型では M5 と M8 とそれ以外の 2 つに，女性型では F3, F5, F7 とそれ以外の 2 つに分かれた。

図 7 は，男性話者，女性話者それぞれの 2 つのクラスタから代表的な 2 名として選択した M5, M13, F2, F7 の声道形状が全話者の平均声道からどのように変位しているかを示す。M5 が含まれるクラスタの話者は咽頭後壁が垂直に近いか後方に傾いており，M13 が含まれるクラスタの話者は咽頭後壁が前方に傾いていた。女性話者でも同様で，F2 が含まれるクラスタの話者は咽頭後壁が垂直に近いかやや後方に傾いており，F7 が含まれるクラスタの話者は咽頭後壁が前方に傾いていた。すなわち，男性話者，女性話者それぞれの 2 つのクラスタは，頭部と頸部の角度の違い，つまり顔を上に向けるか下に向けるかによって分類されていると考えられる。

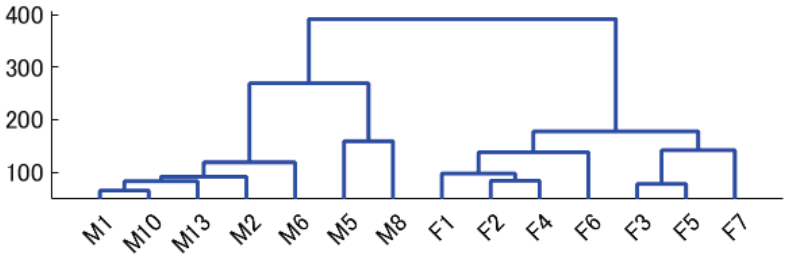

図 6 クラスタ分析の結果

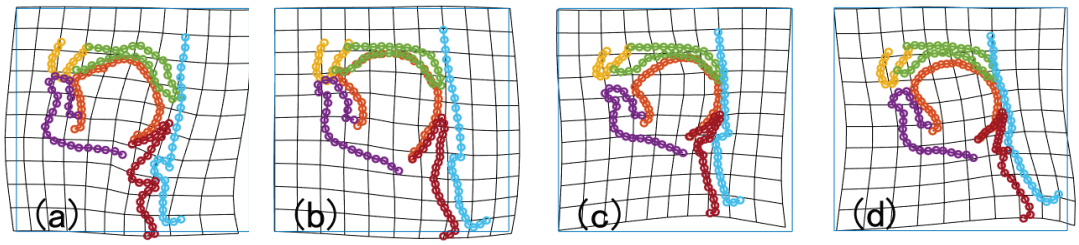

図 7 全話者の平均声道からの変位。(a) M5, (b) M13, (c) F2, (d) F7。

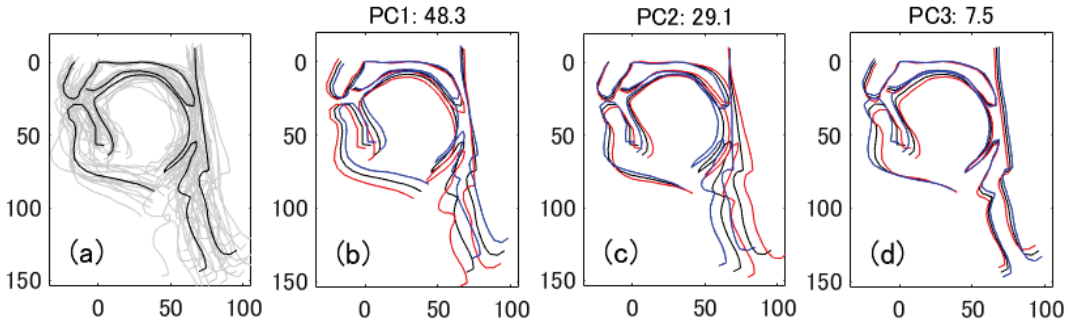

図 8 (a)14 人の輪郭（灰）と平均輪郭（黒），(b) PC1，(c) PC2，(d) PC3。
赤：平均声道＋1 標準偏差，青：平均声道−1 標準偏差。(a)～(d)上部の数値は寄与率。

図 8 は，主成分分析の結果を示し，第 1, 2, 3 主成分（PC1, PC2, PC3）による声道形状の変化と寄与率を示す。PC3 までの累積寄与率は 84.9%であった。PC1 は主として声道の

大きさ，PC2 は主として頭部と頸部の角度，PC3 は主として下顎の前後方向の大きさと口の開きを表していると思われる。これらの成分のうち，PC1 の声道の大きさや PC3 の下顎の前後方向の大きさは形態の個人差や性差を表している。しかし，PC2 の頭部と頸部の角度は，動画撮像が仰臥位で行われたため，発話しやすい首の角度の個人差を表していると考えられる。つまり，顔を上に向けた（喉を伸ばした）方が発話しやすい話者も，顔を舌に向けた（顎を引いた）方が発話しやすい話者もおり，これは性別とは無関係の個人差であると思われる。また，安静呼吸時の口の構えを特に指定していないため，上唇・下唇が接触しているかどうかも性別と無関係の個人差であり，これが PC3 に含まれていると思われる。

図 9 は話者ごとの PC1 と PC2 のスコアである。PC1 は総じて男性話者で大きく，女性話者で小さいことから声道の大きさであり，この成分が図 6 で男性話者と女性話者のクラスタを分ける要因である。PC2 は頭部と頸部の角度を表しており，咽頭壁が垂直に近いかやや後方に倒れている F1, F2, F4, F6 および M5, M8 でスコアが小さく，それ以外では大きい。この成分が男性話者，女性話者内でそれぞれ 2 つのクラスタを分ける要因である。

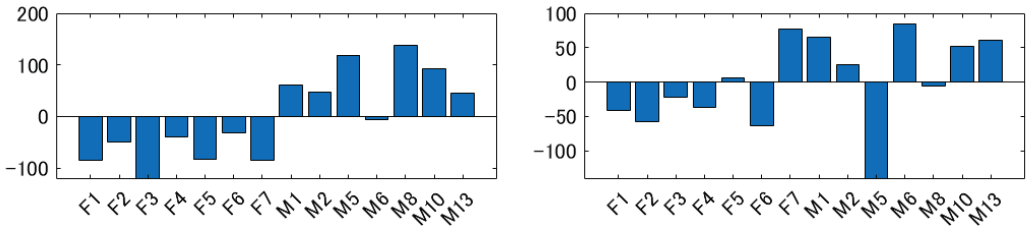

図 9 話者ごとの PC1（左），PC2（右）のスコア。

4．おわりに

本研究では，機械学習を用いて日本語話者 20 名の rtMRI 動画から調音器官の輪郭の抽出を行った。抽出精度の評価を行った結果，喉頭蓋・声帯の部位で精度が低かったが，それ以外の部位では総じて精度が高かった。本研究では，精度評価を各被験者につき 1 本の動画だけで行ったが，同じ被験者の動画であれば評価した動画と同程度の精度で調音器官の輪郭を抽出できる（Takemoto et al. 2019）。よって，20 名の rtMRI 動画からの輪郭抽出は本研究の結果と同程度の精度を持つと考えられる。

輪郭を抽出した 20 名の話者から男女 7 名ずつを選出して声道形状の分析を行った。全話者の平均声道に比べて，男性話者の平均声道は特に上下方向に大きく，喉頭が相対的に下方に位置した。女性話者の平均声道は男性話者と逆に上下方向に小さく，喉頭が相対的に上方に位置した。主成分分析の結果，PC1 は主に声道の大きさ，PC2 は主に首の角度，PC3 は主に下顎の前後方向の大きさや口の開きを表すことが明らかになった。そして，クラスタ分析において PC1 が男女を，PC2 が男性話者および女性話者のクラスタ内での 2 つのクラスタ，すなわち，顔を上に向ける話者と下に向ける話者を分ける主要因であることが明らかになった。つまり，仰臥位における MRI 実験で得られた声道形状には，形態の個人差・性差以外に，発話しやすい首の角度という個人差が大きな要因として含まれていることに留意する必要があることが明らかになった。また，本研究では 16 個の解剖学的なランドマークを設定して安静呼吸時の声道形状の分析を行ったが，発話中の声道形状の分析を行うためには，さらに舌尖や上唇・下唇にもランドマークを置くなどの改良も必要であると考えられる。

謝　辞

本研究は JSPS 科研費 20H01265 の助成を受けて実施した。また, 本研究の遂行にあたり, 輪郭のトレースに取り組んだ宮川翔多氏，オーバーラップの解消のプログラムを開発した並木崇宏氏, 機械学習の導入，トレーサーの開発および輪郭抽出システムを構築した後藤翼氏に感謝する。

文　献

Fitch, W. T., & Giedd, J. (1999). “Morphology and development of the human vocal tract: A study using magnetic resonance imaging,” The Journal of the Acoustical Society of America, 106:3, pp. 1511-1522.

King, D. E. (2009). “Dlib-ml: A Machine Learning Toolkit,” Journal of Machine Learning Research 10, pp. 1755–1758.

Swiderski, Donald L. (1993). “Morphological evolution of the scapula in tree squirrels, chipmunks, and ground squirrels (Sciuridae): an analysis using thin‐plate splines,” Evolution 47:6, pp. 1854-1873.

Takemoto, Hironori, Tsubasa Goto, Yuya Hagihara, Sayaka Hamanaka, Tatsuya Kitamura, Yukiko Nota & Kikuo Maekawa (2019). “Speech organ contour extraction using real-time MRI and machine learning method,” Proc. INTERSPEECH 2019, Graz, 904-908. DOI: 10.21437/Interspeech.2019-1593

後藤翼・荻原裕也・濱中彩夏・竹本浩典・北村達也・能田由紀子・前川喜久雄(2019).「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」日本音響学会 2019 年春期研究発表会講演論文集, 822-823.

後藤翼・竹本浩典・北村達也・能田由紀子・前川喜久雄(2020a).「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」日本音響学会 2020 年春季研究発表会講演論文集, 779-780.

後藤翼・天野沢海・竹本浩典・北村達也・能田由紀子・前川喜久雄(2020b).「rtMRI 動画から抽出した発話器官の輪郭データに基づく平均声道の生成と分析」日本音響学会 2020 年秋季研究発表会講演論文集, 821-822.

前川喜久雄・能田由紀子・北村達也・竹本浩典・石本祐一(2018).「日本語撥音の調音音声学的記述の精緻化：rtMRI データによる試み」日本音響学会 2018 年春季研究発表会講演論文集,1247-1248.

前川喜久雄・西川賢哉・浅井拓也・能田由紀子・正木信夫・島田育廣・竹本浩典・北村達也・斎藤純男・籠宮隆之・石本祐一・菊池英明・藤本雅子・八木豊(2020).「リアルタイム MRI 動画日本語調音運動データベースの設計」言語資源活用ワークショップ 2020 発表論文集, 国立国語研究所コーパス開発センター, 209-230. DOI/10.15084/00003139