日本語話者20名のrtMRI動画における発話器官の輪郭抽出*

☆天野沢海，並木崇宏，宮川翔多，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

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，分析する必要がある[1]。定量的な分析を行うためにはrtMRI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必要がある。

2018年度には，機械学習の導入により，1名の被験者の55本の動画（28,160フレーム）から人間と同程度の精度で発話器官の輪郭を抽出する手法を確立した[2]。2019年度には，18名の被験者の舌の概形をクラスタ分析して特徴的な8名から学習器を生成すれば，全ての被験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を示した[3]。

本研究では，2019年度の研究手法を発話器官の5つの部位に適用し，20名の被験者のrtMRI動画から輪郭を抽出して精度を検討した結果を報告する。

2　材料と方法

2.1 被験者とrtMRI動画撮像

被験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。各被験者はキャリア文「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

各被験者の34本の動画から，同一の語群を含む動画を1本選択して発話器官の輪郭を抽出した。抽出した部位は，口唇・下顎，舌，軟・硬口蓋，咽頭後壁，喉頭蓋・声帯とした。

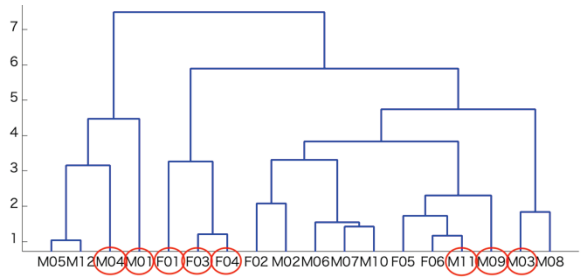

Fig. 1 咽頭後壁の輪郭を抽出する矩形領域のクラスタ分析の結果（赤丸：選出した被験者）

2.3 学習器の生成

まず，20名の被験者から，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被験者を8名選出した（Fig. 1）。次に，選出した被験者の動画から，Table 1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。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Dlib[4]を使用した。

Table 1. トレースするフレーム

部位＼使用枚数	各1枚	各2枚	各3枚	総枚数
口唇・下顎	/a/, /u/, /e/, /m/, 無発話, 第1・512フレーム	/i/, /p/, /o/		13
舌	/a/, /e/, /k/, /t/, /r/, /n/, 無発話, 第1・512フレーム	/i/, /o/, /s/		15
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	/a/, /k/, 無発話	/u/	14

* Speech organ contour extraction from rtMRI videos for 20 Japanese subjects, by AMANO, Takumi, NAMIKI, Takahiro, MIYAGAWA, Shota, GOTO, Tsubasa, TAKEMOTO, Hironori (Chiba Institute of Technology), KITAMURA, Tatsuya (Konan Univ.), NOTA, Yukiko, and MAEKAWA, Kikuo (NINJAL).

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[2]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。

3　結果と考察

Table 2 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[2]。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[2]。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。

Table 2 輪郭抽出の誤差（単位：pixel）

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
口唇･下顎	0.96	0.75	0.56	0.69	**1.23**	-	-	**1.40**	-	0.71
舌	0.96	0.90	-	-	-	0.78	**1.78**	**1.29**	0.68	0.71
軟･硬口蓋	0.68	**1.51**	-	-	-	0.80	-	0.67	0.94	0.60
咽頭後壁	-	**1.12**	-	-	**1.21**	**1.18**	**1.45**	0.96	-	0.66
喉頭蓋･声帯	**1.22**	-	-	0.95	**1.42**	0.88	**2.02**	**1.93**	**1.24**	**1.50**

	M11	M12	M13	F1	F2	F3	F4	F5	F6	F7
口唇･下顎	**1.15**	-	0.50	-	-	0.72	**1.02**	-	-	0.99
舌	-	-	0.55	-	**1.13**	0.67	-	-	**1.13**	**1.12**
軟･硬口蓋	0.60	-	0.75	-	**1.10**	**1.15**	0.49	-	-	0.70
咽頭後壁	-	**1.56**	**1.19**	-	**1.07**	-	-	0.87	0.87	0.70
喉頭蓋･声帯	-	-	**1.03**	**1.27**	-	-	-	**1.75**	-	0.53

Table 3 輪郭抽出の精度の 5 段階評価

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
口唇･下顎	**4**	4	4	4	3	-	-	3	-	5
舌	4	4	-	-	-	4	5	4	3	**4**
軟･硬口蓋	3	**2**	-	-	-	5	-	5	3	5
咽頭後壁	-	4	-	-	**4**	5	4	**4**	-	4
喉頭蓋･声帯	5	-	-	5	**2**	4	5	**2**	5	5

	M11	M12	M13	F1	F2	F3	F4	F5	F6	F7
口唇･下顎	3	-	4	-	-	4	4	-	-	4
舌	-	-	4	-	4	3	-	-	4	3
軟･硬口蓋	4	-	3	-	3	4	5	-	-	**2**
咽頭後壁	-	5	**2**	-	5	-	-	4	5	**2**
喉頭蓋･声帯	-	-	4	**1**	-	-	-	**1**	-	3

Table 3 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 Table 2 で誤差が小さかった動画は Table 3 で評価が高かった。一方，Table 2 で誤差が大きかった動画は，必ずしも Table 3 で評価が低いとは限らなかった。例えば，Table 2 で M7 の喉頭蓋･声帯の誤差は 2.02 で最大であったが，Table 3 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。

本研究では，部位ごとに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（Fig. 2）。

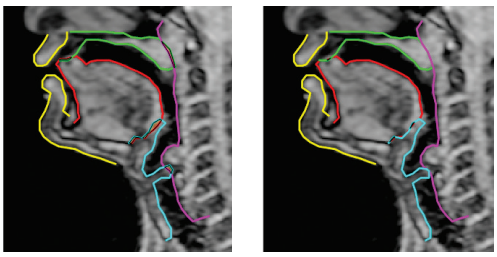

Fig. 2 オーバーラップ解消処理（左：処理前，右：処理後）

4　まとめ

本研究では，日本語話者 20 名の rtMRI 動画から発話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発話器官の輪郭を抽出できる[2]。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。

謝辞

本研究は JSPS 科研費 20H01265 の助成を受けた。

参考文献

[1] 前川ら，音講論（春），1247-1248，2018.
[2] Takemoto *et al.*, Proc. of Interspeech 2019, 904-908, 2019.
[3] 後藤ら，音講論（春），779-780，2020.
[4] King, Mach. Learn. Res., 1755-1758, 2009.