F0推定に寄与する声道の輪郭成分の検討*

◎脇田真子，竹本浩典（千葉工大），平井啓之，前川喜久雄（国語研）

1　はじめに

ソース・フィルタ理論では，抑揚の要因となる基本周波数（F0）は声帯の緊張や肺圧で制御され，声道形状の制御とは独立とされる。しかし，大谷らは声道形状が記録されている1枚のMRI正中矢状断面画像からF0を推定することができることを示した[1]。さらに，われわれは前報でMRI画像より情報の少ない声道の輪郭から音響特徴量を予測し，ボコーダーを用いて音声合成を行い，音韻性や抑揚を再現した合成が可能であることを示した[2]。これらはソース・フィルタ理論に反する結果と言える。本研究では，声道の輪郭からF0を予測するモデルを再構築し，説明可能なAI（Explainable AI: XAI）を用いて，F0推定に寄与する声道の輪郭成分を検討する。

2　材料と方法

2.1 MRI実験・録音

実験参加者は日本人男性1名（60代）である。実験参加者はMRI内で仰臥し，ATR503文[3]を2日に分けて読み上げた。その調音運動をリアルタイムMRI（rtMRI）で動画として記録し，同時にOptoacoustics製FOMRIタイプ光マイクロホンを用いて音声も収録した。撮像はATR-Promotionsに設置されているSiemens製MAGNETOM Prisma fit 3Tで行い，フレームレートは27 fps, 解像度は1×1 mm, スライス厚は10 mm，画像サイズは256×256 pixelであった。

2.2 データセット

2.1で得られたデータから実験で用いるためのデータセットを作成した。まず，2段階抽出法[4]を用いてrtMRI動画の各フレームからFig. 1で示す調音器官の5部位の輪郭をpixel座標の点群として自動抽出した。なお，輪郭点は舌40点，口唇・下顎40点，軟・硬口蓋30点，咽頭後壁・披裂部28点，喉頭蓋・声帯30点であり，168（輪郭点数）×2（各点のx，y座標）の336次元のベクトルである。これを声道の輪郭とした。また，MRI撮像は2日間にわたって行われたため，撮像日によって画像内の体の位置が変動する。そのため，1日目のデータと2日目のデータそれぞれから軟・硬口蓋の前鼻棘，咽頭後壁の最上部と最下部の3点が一致する剛体変換行列を求めて位置合わせした。

次に，MRI撮像と同時に録音した音声データからF0を抽出した。まず，音声データからスペクトルサブトラクション法を用いてMRI装置の駆動音を除去した。そして，この音声からWORLD [5] (D4C edition [6])を用いてフレームシフト5 msでF0を抽出し，rtMRI動画のフレームレートに合わせてフレームシフト37 msにダウンサンプリングした。

以上の処理で得られた声道の輪郭およびF0は，それぞれのデータの平均が0，分散が1になるように正規化し，有声区間のフレームのみを用いてデータセットとした。2日間の撮像で得られた503文章の発話のうち，403文章を訓練データ，50文章を検証データ，50文章をテストデータとした。

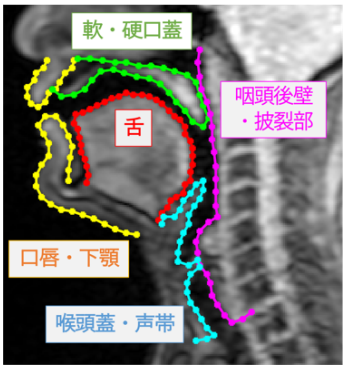

Fig. 1 調音器官の5部位

2.3 F0推定モデル

本研究では，全結合層を用いて声道の輪郭からF0を推定するモデルを作成した。このモデルは入力として1フレームの声道の輪郭データを受け取り，6つの全結合層を通して

* Investigation of contour components of the vocal tract contributing to F0 estimation, by WAKITA, Mako, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (NINJAL).

1 次元の F0 を出力する。各中間層次元数は 2048 とした。損失関数に平均二乗誤差，最適化手法に学習率 0.0001 の Adam を用いた。

2.4 F0 推定に寄与する輪郭成分の可視化

F0 推定に寄与する声道の輪郭成分を可視化する手法として Integrated Gradients [7]を用いた。これは，モデルの出力に対する各入力特徴の影響度を計算する手法である。基準となる baseline の形状から入力値となる input の形状まで，対応する輪郭点を徐々に変位させる過程で計算された F0 の勾配の平均値を求め，変位量を乗じることで，形状の変位が F0 に与える影響度を算出する。

本研究では，各母音の平均声道形状を baseline とし，各母音を 120, 200 Hz 近辺で発話しているフレームから作成した平均声道形状を input とした。各母音の平均 F0 が約 160 Hz であるため，それぞれ F0 上昇・下降に寄与する輪郭成分を可視化した。

3 結果と考察

3.1 F0 推定結果

1 フレームの声道の輪郭から推定した F0 の二乗平均平方根誤差（Root Mean Squared Error: RMSE）は 23.26 Hz であった。前報では 5 フレームの声道の輪郭から推定した F0 の RMSE が 28.53 Hz であったが，フレーム数を削減したにもかかわらず精度が向上した。これは学習に有声区間のフレームのみを用いたため，より F0 推定に有効な情報のみを学習させたためと考えられる。

3.2 F0 推定に寄与する成分の可視化結果

Fig. 2 は F0 上昇，下降に寄与する輪郭成分を示す。灰色の線が平均声道形状の輪郭（baseline），黒の破線が各母音を 120, 200 Hz 近傍で発話した平均声道形状の輪郭（input）である。矢印は変位の方向で，青いほど F0 が下降し，赤いほど F0 が上昇する。

F0 上昇時，下降時は喉頭がそれぞれ前上方，後下方に移動した。また，/a/や/o/では舌の変位も見られたが，矢印の色が喉頭蓋で濃いことから，F0 推定モデルではこの部分の輪郭の変位が F0 に対して大きな影響を持つ。F0 に作用する調音運動として，舌骨を前方に牽引して声帯を伸張させる動作や，喉頭を下降させる動作がある[8]。これらが喉頭蓋部分の輪郭に影響を与えていると考えられる。

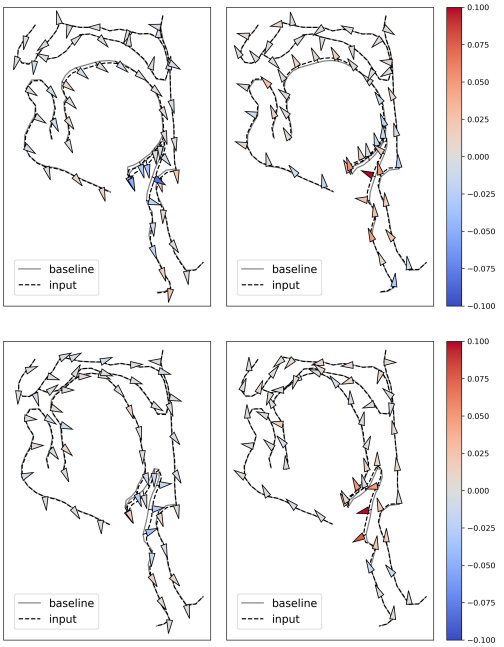

Fig. 2 F0 上昇・下降に寄与する輪郭成分の可視化結果（左上: /a/ 120 Hz，右上: /a/ 200Hz，左下: /i/ 120 Hz，右下: /i/ 200 Hz）

4 まとめ

本研究では，XAI 技術により F0 推定モデルにおいて影響の大きい声道の輪郭成分を可視化した。その結果, F0 上昇・下降の双方に，喉頭蓋の影響が大きいことが明らかになった。

今後は，他の話者でも同様の分析を行い，本研究で得られた結果を検証する。また，テキストから声道形状を予測するモデル[9]を用いて，アクセントを変えた際の声道形状の変動についても本研究と同様の結果が得られるか検証する。

謝辞

本研究は JSPS 科研費 24K00071 の助成を受けて実施した。

参考文献

[1] 大谷ら,音講論（秋），1117-1118, 2023
[2] 脇田ら,音講論（春），1299-1300, 2024.
[3] 小林ら,音響学会誌, 48(12),888-893,1992.
[4] 藤澤ら,音講論（秋），1015-1016, 2022.
[5] Morise *et al.*, IEICE, pp. 1877-1884, 2007
[6] Morise, Speech Communication, vol. 84, pp.57-65, 2016
[7] Sundararajan *et al.*, *Proc. ICML*, 3319-3328, 2017.
[8] 本多清志, *実験音声科学-音声事象の成立過程を探る-*, コロナ社，2018.
[9] 脇田ら,音講論（秋），927-928, 2023