リアルタイムMRI動画から抽出した声道の輪郭に基づく音声合成*

☆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

調音モデルから音声を合成する技術は調音運動と音声の関連の解明に寄与する。近年では深層学習の発展に伴い，EMA（Electro Magnetic Articulography）やリアルタイムMRI（rtMRI）で観測した調音運動情報から深層学習モデルを用いて音声を合成する手法が多く研究されている[1-3]。

EMAは口唇や舌などにセンサコイルを装着し，調音運動による変位を記録する技術である。そのため，変位と音響特徴量の関係を容易に検討できる。しかし，センサコイルの設置できる箇所が少数であるため，観測できる調音運動は限られている。一方，rtMRIは調音器官の全体的な運動を撮像できるが，調音器官各部の変位を得るために，各フレームで調音器官の輪郭を抽出する必要がある。

本研究では，rtMRIデータから調音器官の輪郭を抽出し，これらの座標データから基本周波数，スペクトル包絡等の音響特徴量を予測し，音声の合成を試みたので報告する。

2　材料と方法

2.1 MRI実験・録音

実験参加者は日本人成人男性1名（60代）である。実験参加者はMRI内で仰臥し，ATR503文[4]を2日に分けて読み上げた。その調音運動を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 3で行い，フレームレート27 fps，解像度1×1 mm，スライス厚10 mm，画像サイズ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の作成を行った。まず，2段階モデル抽出法[5]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。

次に，MRI撮像と同時に録音した音声データから音響特徴量の抽出を行った。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[6]を用いてフレームシフト5 msで1次元の基本周波数（F0），45次元のメル一般化ケプストラム（mel-generalized cepstrum: mgc），2次元の非周期性指標（coarse aperiodicity: cap）を抽出し，rtMRI動画のフレームレートに合わせて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音響特徴量の全てのパラメータは，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，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。

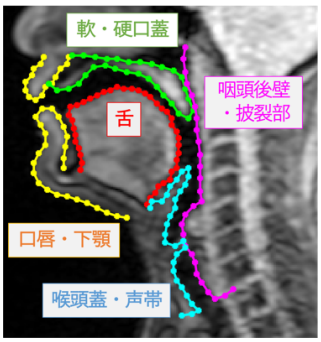

Fig. 1 調音器官の5部位

2.3 調音運動予測モデル

本研究では，全結合層を用いて声道の輪郭から音響特徴量を予測するモデルを提案する。このモデルは入力として当該フレーム＋前後各2フレームの計5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各

* Speech synthesis based on vocal tract contours extracted from real-time MRI videos, by WAKITA, Mako, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (NINJAL).

音響特徴量を出力する。各中間層の次元数は2048，出力次元数は各音響特徴量の次元数とした。損失関数には平均二乗誤差，最適化手法には学習率 0.0001 の Adam を用いた。

2.4 予測精度の評価

本モデルの予測精度を評価するために客観評価実験と主観評価実験を行った。

客観評価実験では，音響特徴量ごとにテストデータ 50 文章での予測値と元音声から抽出した値との誤差を算出した。誤差算出方法は，F0 と cap は二乗平均平方根誤差（Root Mean Squared Error: RMSE），mgc は mgc のみ予測した値を用いて合成した音声と再合成した元音声からそれぞれ SPTK[7]を用いて求めた 25 次元のケプストラム距離を採用した。

主観評価実験では，聴覚に問題のない 16 名がテストデータから無作為に抽出した 10 文章に対して抑揚の再現度，音韻性，品質を 1~5 で 5 段階評価した。抑揚と品質の評価実験では，元音声から再合成した音声を基準とし，予測した F0，mgc，cap（抑揚評価時は F0 のみ予測値）を用いて合成した音声との比較を行った。音韻性の評価実験では，mgc のみ予測値を用いた合成音声が目標とする文の音韻情報を再現できているか評価した。なお，rtMRI の時間分解能は 27 fps だが，音声合成ではより高い時間分解能が必要である。そこで，音響特徴量をフレームシフト 5 ms になるように線形補間して WORLD で合成した。

3　結果と考察

3.1 結果

Table 1 にテストデータでの客観評価実験の結果を示す。F0 の RMSE は 28.53 Hz，ケプストラム距離は 7.36 dB，cap の RMSE は 1.30 であった。また，客観評価実験の結果を Fig. 2 に示す。抑揚の再現度は 3.69，音韻性は 4.16，音声の品質は 3.31 であった。

3.2 考察

基本周波数は声帯の緊張や肺圧で決定されるため，正中矢状断面の声道の輪郭からこれらの情報を十分に得ることは困難であると予想された。しかし，抑揚の評価値（3.69）は 3 の評価値「部分的に抑揚を再現できていないところがあり，わずかに気になる」より値が高かった。そのため，声道の輪郭からかなりの精度で F0 を推定できたといえる。また，声道形状に由来する音韻性の評価値（4.16）は，音韻性は十分に再現できたことを示す。音声の品質（3.31）は，音声としては十分に聞き取れるが元音声と比べると劣化が認められるという値である。これは抑揚や音韻性の劣化に加え，F0 と mgc の整合性やフレーム間の連続性の考慮が出来ていないことが原因ではないかと推測される。

Table 1 各音響特徴量の客観評価値

評価種目	評価値（標準偏差）
F0 RMSE [Hz]	28.53 （5.37）
ケプストラム距離 [dB]	7.36 （0.77）
cap RMSE	1.30 （0.16）

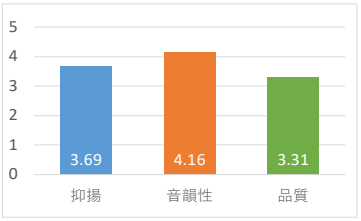

Fig. 2 主観評価結果

4　まとめ

本研究では，rtMRI 動画から抽出した声道の輪郭から音響特徴量を予測し，WORLD ボコーダーを用いて音声の合成を試みた。音声の品質は元音声と比べると劣化が認められたが，声道の輪郭から音韻性や抑揚を再現した合成が可能という結果が得られた。

今後は品質の改善のため，全結合層モデル以外のモデルの検討や声道の輪郭座標点が及ぼす音響特徴量への影響の検討を行う。

謝辞

本研究は JSPS 科研費 20H01265 の助成を受けて実施した。

参考文献

[1] Wu *et al.*, Proc Interspeech2022, 779-783.
[2] Otani *et al.*, Proc.Interspeech2023, 127-131.
[3] Wu *et al.*, Proc Interspeech2023, 5132-5136
[4] 小林ら，音響学会誌, 48 (12),888-893, 1992.
[5] 藤澤ら，音講論　(秋)，1015-1016, 2022.
[6] Morise *et al.*, IEICE, pp. 1877-1884, 2007
[7] SOURCEFORGE.NET, 2021,”Speech Signal Processing Toolkit (SPTK)”,http://sp-tk.sourceforge.net/, (参照 2023/12/28).