リアルタイムMRIから抽出した声道形状を介したテキスト音声合成*

☆新飼遥彩，△安田彩乃，△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

近年，リアルタイム MRI（rtMRI）で撮像した調音器官の形態や調音運動情報から音声を合成する研究が行われている[1]。大谷らは rtMRI のビットマップ情報から音声を合成した[2]。脇田らは，rtMRI のビットマップから抽出した調音器官の輪郭を用いて言語特徴量から声道形状を予測するモデル（モデル①）[3]と声道形状から基本周波数やスペクトル包絡等の音響特徴量を予測するモデル（モデル②）[4]を構築した。モデル①では声道形状の輪郭点の平均誤差が 1.19 pixel の精度で予測できた。モデル②では合成音声の主観評価値が3.31となり元音声から再合成した基準音声と比較して劣化が確認されたが，ある程度の再現性を示した。しかし，モデル①，②を直列結合した場合，すなわちモデル①＋②の場合では，言語特徴量から音声合成すると評価値は 2.55 で基準音声から大幅に劣化した。そこで本研究では，これを改善するためにモデル①と②を統合して音声の合成を試みた。

2　材料・方法

2.1 MRI 撮像・録音

日本人成人男性 1 名（60 代）が ATR503 文[5]を 2 日に分けて読み上げた調音運動を rtMRI で動画として記録し，同時に光マイクロホンを用いて音声も収録した。これらは[3,4]と同一のデータである。

2.2 データセット

まず，MRI 撮像中に録音された音声からスペクトルサブトラクション法を用いて MRI 装置のノイズを除去した。次に，音声波形に音素ラベリングを行い，言語特徴量を作成した[6]。これは，HTS のフルコンテキストラベルから生成される当該音素や隣接音素，アクセントなどの特徴量に継続時間長，また撮像日の違いを考慮するために日付 ID を加えたもので，動画のフレームごとに 294 次元のベクトルである。

また，rtMRI 動画の各フレームから 2 段階抽出法[7]を用いて発話器官の 5 部位（舌，口唇・下顎，軟・硬口蓋，咽頭後壁・披裂部，喉頭蓋・声帯）の輪郭を 168 点で抽出した。各点は x, y 座標からなるため，フレームごとに 168×2＝336 次元のベクトルである。

そして，発話音声に対して WORLD[8]を用いて音響特徴量を抽出した。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）を得た。また，有声区間と無声区間を判別するため，2 値の有声無声フラグ（voiced/unvoiced flag: V/UV）を設定した。これらを rtMRI 動画のフレームレート 27.17 fps に合わせてダウンサンプリングし，フレームごとに 49 次元のベクトルとした。

以上の V/UV を除く全てのデータは，それぞれ平均が 0，分散が 1 となるよう正規化した。2 日間の実験で得られた計 503 文章の発話のうち 402 文章を訓練，50 文章を検証，51 文章をテストデータにランダムに割り当てた。

2.3 声道形状・音響特徴量予測モデル

本研究では，言語特徴量を入力すると声道形状と音響特徴量を予測し，声道形状を入力すると音響特徴量を予測する統合モデルを構築した。まず，モデル①と②を従来手法[3,4]に基づき個別に学習した後，両モデルを連結し，学習済のパラメータを初期値としてモデル全体を再学習した。再学習では，言語特徴量を入力とし，声道形状と音響特徴量の予測誤差を重み付きで加算した損失関数を最小化した。学習率を低く設定し，各誤差の重みは試行錯誤により決定した。なお，モデル①は

* Text-to-speech synthesis via vocal tract shape extracted from real-time MRI, by SHINKAI, Sumisa, YASUDA, Ayano, WAKITA, Mako, TAKEMOTO, Hironori (Chiba Institute of Technology), HIRAI, Hiroyuki, and MAEKAWA, Kikuo (National Institute for Japanese Language and Linguistics).

1 層の全結合層，モデル②は 1 層の畳み込み層と 6 層の全結合層を用いた。学習パラメータは，中間素子数 5120，損失関数は平均二乗誤差，活性化関数は ReLU，最適化手法は学習率 1e-4（再学習時 1e-6）の Adam とした。ドロップアウト率はモデル①で 0.1, モデル②で 0.4 とした。

2.4 予測精度の評価

統合モデルの予測精度を評価するため，テストデータに対して客観評価実験と主観評価実験を行い，モデル①，②と比較した。声道形状予測では，予測した輪郭点座標と抽出した輪郭点座標との二乗平均平方根誤差（Root Mean Square Error: RMSE）を算出した。音響特徴量予測では，F0 と cap は予測値と元音声から抽出した値との RMSE，mgc は mgc のみ予測値を用いて合成した音声と元音声から再合成した音声とのケプストラム距離を SPTK[9]を用いて算出した。V/UV は判定の誤り率を算出した。

主観評価実験は，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 文章に対して音質を 1~5 の 5 段階で DMOS 評価した。発話音声から WORLD を用いて分析再合成した音声を基準とし，言語特徴量を入力して予測した音響特徴量を用いて合成した音声と比較した。なお予測は，全ての音響特徴量（all），F0 のみ（F0），mgc のみ（mgc），cap のみ（cap）の 4 種類とした。

3　結果と考察

Table 1 に声道形状予測の結果を示す。統合モデルの予測精度はモデル①とほぼ同等であった。Table 2 に音響特徴量予測の結果を示す。抽出した声道形状を入力した場合，統合モデルの予測精度は mgc を除く全音響特徴量で同等または有意に高かった。予測した声道形状を入力した場合，統合モデルの予測精度は F0 を除く全音響特徴量で有意に高かった。

Fig. 1 に合成音声の主観評価の結果を示す。統合モデルでの all の評価値は 3.31 で，基準音声に比べて「劣化がわずかに気になる」より少し高い評価であった。all では，モデル①＋②より有意に高い評価値が得られた。一方，単一の予測値を用いた音声の評価値は 3.79～4.83 に分布し，いずれもモデル①＋②と同等かやや上回った。単一の予測値を用いた音声が高評価であったことから，予測した声道形状からでも各音響特徴量は一定の精度で予測可能だといえる。

Table 1　声道形状の客観評価値（単位：pixel）

発話器官	統合モデル	モデル①
舌	1.85±0.57	1.84±0.56
口唇	0.89±0.33	0.8±0.33
軟・硬口蓋	0.70±0.30	0.70±0.30
咽頭後壁	0.74±0.21	0.74±0.21
喉頭蓋・声帯	1.49±0.44	1.49±0.44
平均	1.14±0.61	1.13±0.61

Table 2　音響特徴量の客観評価値

評価項目	抽出した声道形状を入力			予測した声道形状を入力		
	統合モデル	モデル②	有意差	統合モデル	モデル①＋②	有意差
F0 [Hz]	22.61±8.29	22.34±8.78	n.s.	29.74±9.02	29.57±6.29	n.s.
mgc [dB]	7.45±0.57	7.37±0.60	***	7.67±0.56	8.03±0.62	***
cap	1.35±0.20	1.35±0.20	n.s.	1.33±0.21	1.37±0.21	***
V/UV [%]	8.35±2.95	14.45±5.04	***	7.38±2.84	14.44±4.64	***

t検定(***p<.001)

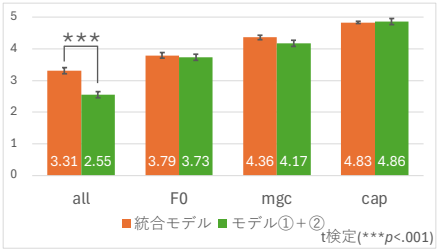

Fig. 1　主観評価の評価値

4　まとめ

本研究では，言語特徴量から統合モデルで声道形状と音響特徴量を予測し，音声を合成した。その結果，多くの特徴量予測においてモデル①，②，①＋②と同等またはそれ以上の精度を示した。合成音声は基準音声よりもわずかに劣化したがモデル①＋②より音質が向上し，単一の音響特徴量予測は再現できた。

今後は，任意テキストや話者拡張に対応した音声合成の実現を目指す。

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] Wu *et al.*, Proc Interspeech2023, 5132-5136, 2023.

[2] Otani *et al.*, Proc. Interspeech2023, 127-131, 2023.

[3] 脇田ら，音講論（秋），927-928, 2023.

[4] 脇田ら，音講論（春），1299-1300, 2024.

[5] 小林ら，音響学会誌, 48 (12) , 888-893, 1992.

[6] Boersma, ,Glot International 5:9/10, 341-345, 2001.

[7] 藤澤ら，音講論（秋），1015-1016, 2022.

[8] Morise *et al.*, ICECE, E99.D (7) , 1877-1884, 2016.

[9] SPTK:音声信号処理ツールキット, http://sp-tk.sourceforge.net/, (参照 2025/7/6).