Wav2lip モデルのファインチューニングによる rtMRI 動画の生成*

☆陸鳴宇 (甲南大院), 北村達也 (甲南大), 前川喜久雄 (国語研)

1　はじめに

音声からその発声時の声道形状を逆推定する問題は，古くから多くの音声研究者の興味を集め，これまでに多数の研究が行われてきた．しかし，この問題は不良設定問題であるため，正確な推定は困難とされてきた．一方，生成系 AI 技術の劇的な進歩により，高品質な画像，動画，音声等の生成が可能になっている．その中には，wav2lip [1] と sadtalker [2] などのリップシンク動画を生成する技術が含まれる．本研究では，この技術を応用して音声から real-time MRI (rtMRI) 動画を生成する手法を提案する．

最近，深層学習にもとづいて音声から調音運動を逆推定する手法が提案されている．Csapó [3] は，CNN-LSTM を含む深層学習モデルを利用し，音響的特徴から rtMRI 動画を生成し，調音運動を推定する手法を提案している．このアプローチは，音声生成プロセスの理解に役立ち，音声認識システムの改善に利用できる可能性がある．国内でも大浦ら [4] は Csapó の手法を踏まえて，LSTM を BLSTM に切り替えることで rtMRI 動画の生成精度を向上させた．また，梶浦ら [5] は音声信号を利用して，音声器官の輪郭形状の推測を行うなど，この課題の研究が活発化している．

本研究では，上記の研究手法とは異なり，音声と顔画像 (動画) を学習して作成されたリップシンク動画生成用の深層学習モデルを活用し，このモデルをファインチューニングすることによって音声から rtMRI 動画を生成する技術を提案する．ファインチューニングには『リアルタイム MRI 日本語調音運動データベース』 [6] の構築のために収集された rtMRI データを利用する．本研究により生成 AI の応用範囲がひろがることを期待している．

2　研究手法

2.1　rtMRI データ収集

日本語母語話者 32 名を対象に日本語の種々の音節および文を発話した際の正中矢状面の rtMRI データを用いた [6]．撮像は，ATR-Promotions 脳活動イメージングセンタの MRI 装置 Siemens Magnetom Prisma 3T にて行った．画像サイズは 256 × 256 ピクセル，解像度は 1 mm × 1 mm，スライス厚は 10 mm，フレームレートは 14 fps または 27 fps である．rtMRI データと同時に光マイクロフォンを用いて標本化周波数 48 kHz，量子化 16 bit にて音声を収録した．音声に重畳された撮像ノイズは Zhao ら [7] のアルゴリズムにて除去した上で，rtMRI 動画に付与した．rtMRI 動画の総時間数は約 15 時間である．

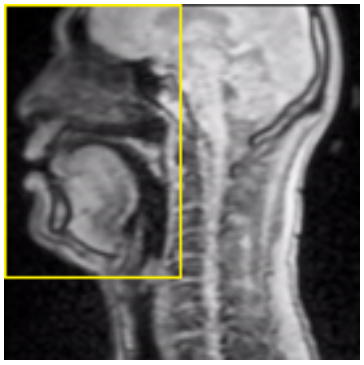

Fig. 1　Original rtMRI and trimmed area(yellow rect) images

Table 1　Dataset.

Set	Number of movies
Train	11,206
Validation	1,401
Test	1,316

上記のデータと別に，中国語母語話者 1 名 (話者 W) を対象に日本語版と中国語版の「北風と太陽」発話時の rtMRI および音声データを収録した．データ収集に関する条件は日本語母語話者と同じである．

2.2　データセット

Fig. 1の黄色領域のように下半分が舌領域となるよう rtMRI 動画をトリミングした上で，フレームレートを 25 fps，画像サイズを 1920 × 1080 ピクセルに変換した．さらに，文献 [1] に合わせて動画を約 5 s ごとに分割した．その結果，日本語母語話者のデータ数を Table 1に示す．

データのグループ分けはランダムに行った．Test セットは評価専用のため，訓練には用いていない．また，話者 W のデータも訓練に用いていない．

2.3　rtMRI 生成モデルと学習

本研究では，リップシンク動画生成プロジェクト wav2lip [1] のネットワーク構造を利用し，ファインチューニングの手法を使用して，rtMRI 動画を生成した．wav2lip は生成器 1 個と識別器 2 個の GAN を使っている (Fig. 2)．

ネットワークに音声と顔動画の 1 フレームを入力

*Generation of rtMRI movies by fine tuning of Wav2lip model. by LU, Mingyu (Konan Univ. Graduate School), KITAMURA, Tatsuya (Konan Univ.) and MAEKAWA, Kikuo (NINJAL)

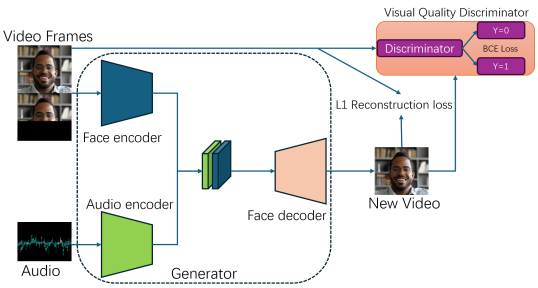

Fig. 2　wav2lip network [1]

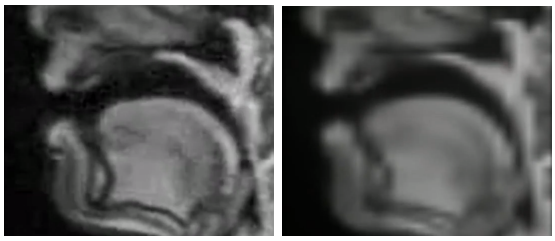

Fig. 3　Example of the for the condition 1. original (left) and generated (right) images

する．それらの入力を基づいて，まず音声から音響的特徴 (MFCC) を抽出し，次に動画フレームから口唇形状を抽出し，最後に抽出したそれらの特徴を組み合わせて口唇の動きの動画を生成する．GitHub 上で公開されているモデルは LRS2 [8] というデータセットを用いて訓練されている．

本研究では wav2lip の訓練済みの checkpoint に編集後の rtMRI データを追加するというファインチューニングを行った．Early Stopping を設置した上で学習を始め，L1loss 値が 0.0025 付近で安定になり訓練を終了した．

2.4　生成精度の評価

訓練済みモデルの効果を確認するため，3 組の検証データを準備した．条件 1 は，Test セットからランダムに得たのも，条件 2，3 は話者 W による「北風と太陽」のそれぞれ日本語と中国語データを使用した．その 3 条件で各 5 つの動画を生成して，オリジナルとの差を Structural Similarity (SSIM) と Normalized Mean Squared Error (NMSE) で評価した．SSIM は 1 に近いほど画像が類似していることを示し，NMSE は値が小さいほど (すなわち 0 に近いほど) 画像が類似していることを示している．生成した動画とオリジナル動画の下半分 (Fig. 3) を取り出し計算を行った．

3　結果と評価

オリジナル動画と生成された動画の SSIM と NMSE を Table 2に示す．すべての条件で SSIM は比較的高い値を示し，NMSE は 0.02 未満であった．この結果は，提案法によってオリジナルに近い動画が生成さ

Table 2　Experimental results.

	Condition		
	1	2	3
SSIM	0.9074	0.8256	0.8154
NMSE	0.0146	0.0187	0.0165

れたことを示している．条件間で比較すると，条件 1 の SSIM が最も高く，NMSE が最も低い．これは，Test セットの話者はモデルの学習にも用いられているためと考えられる．一方で，条件 2 と条件 3 の結果は，モデルの学習に用いられなかった話者の生成精度は用いた話者よりも悪化することを示している．条件 2 と条件 3 の SSIM と NMSE の差は小さく，生成精度の言語依存性が小さいことを示している．ただし，Table 2の数値は平均値であり，日本語には存在しない中国語の音に対応する動画の精度については，今後，詳細に分析する必要がある．

4　おわりに

本研究では，リップシンク動画生成の wav2lip ネットワークにファインチューニングを行い，rtMRI 動画の生成を試みた．今後の課題として，生成した動画の質評価する方法を検討することが挙げられる．また，今回の学習に使用したのは日本語データだけであったが, 他の言語のデータの追加することも検討する．

謝辞　本研究の一部は，JSPS 科研費 (No. 24K00071) および甲南デジタルツイン研究所の支援を受けた．

参考文献

[1] Prajwal *et al.*, *Proc. 28th ACM International Conference on Multimedia*, 484–492 (2020).
[2] Zhang *et al.*, *Proc. IEEE/CVF Conf. on CVPR*, 8652-8661 (2022).
[3] Csapó, *Proc. INTERSPEECH 2020*, 3720-3724 (2020).
[4] 大浦ら, 音講論 (春), ROMBUNNO.3, 27 (2024).
[5] 梶浦ら, 音講論 (春), 1285–1286 (2024).
[6] 前川ら, 言語資源活用クークショップ 2020, 209-230 (2020).
[7] Zhao *et al.*, *Proc. ICASSP 2022*, 9281-9285 (2022).
[8] The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset. `https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html`