1-4-20

Wav2lip モデルのファインチューニングによる rtMRI 動画の生成

Generation of rtMRI movies by fine tuning of Wav2lip model

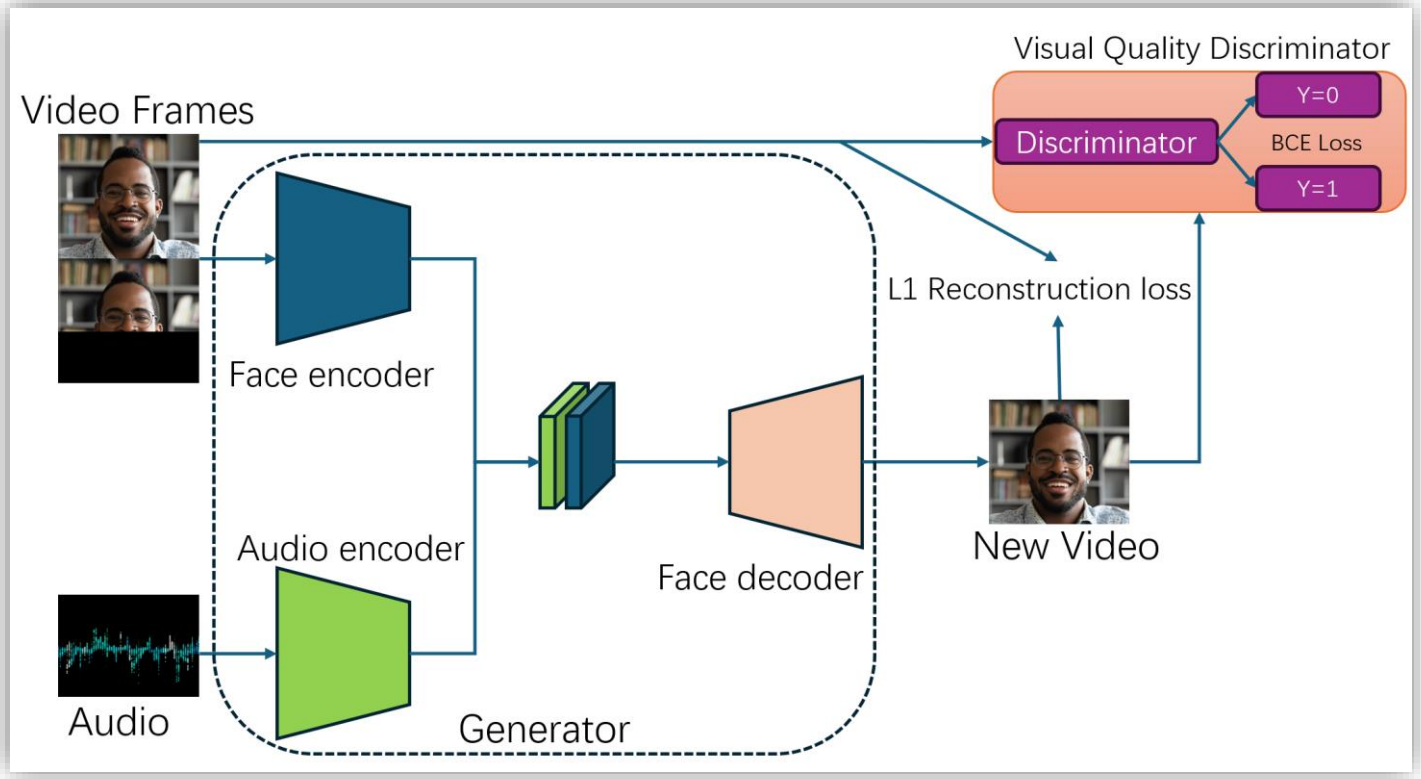
☆陸鳴宇(甲南大院), 北村達也(甲南大), 前川喜久雄(国語研)

研究の背景と目的

- 音声からその発声時の声道形状を逆推定する問題は，これまでに多数の研究が行われてきたが．正確な推定が難しいとされてきた．
- 近年では，生成系AI技術により，高品質な音声や動画などの生成が可能となった．深層学習に基づく調音運動の逆推定法の研究も多数行われている（Csapó ,2020; Zhang et al., 2022; 大浦ら, 2024; 梶浦ら, 2024など）．
- 本研究では，リップシンク動画生成用の深層学習モデル[1]を活用し，音声からrtMRI 動画を生成する技術を提案．

[1] Prajwal et al., Proc. 28th ACM International Conference on Multimedia, 484–492 (2020).

Wav2lip(Prajwal et al., 2020)



音声と顔動画1フレームごと入力

↓

MFCC抽出

↓

口唇形状抽出

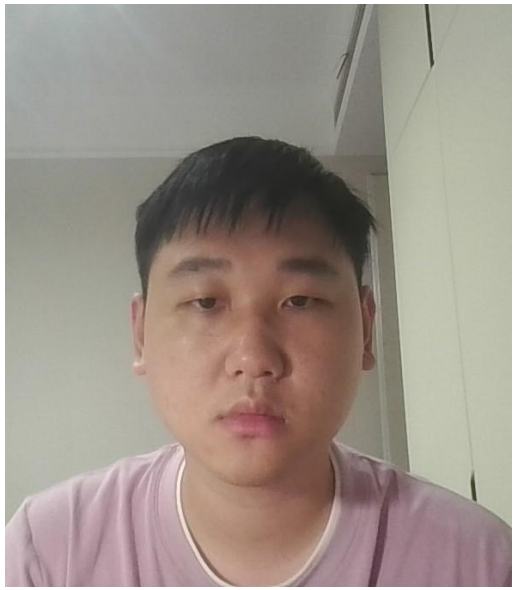
↓

リップシンク動画

[1] Prajwal et al., Proc. 28th ACM International Conference on Multimedia, 484–492 (2020).

Wav2lipの実行例

ネットワーク入力



顔写真

音声（任意言語）

ネットワーク出力

顔動画

研究に用いたrtMRIデータ(1/2)

日本語母語話者データ(学習・評価用)

32名が日本語の種々の音節および文を発話 [2].

- rtMRIはSiemens Magnetom Prisma 3Tで撮像
- 音声は光マイクロフォンで収録

中国語母語話者データ(評価用)

1 名（話者 W）が**日本語**版と**中国語**版の「北風と太陽」を発話.

- rtMRIはSiemens Magnetom verioで撮像
- 動画の総時間約15時間

[2] Maekawa, Acoustical Science and Technology, (2024).

研究に用いたrtMRIデータ(2/2)

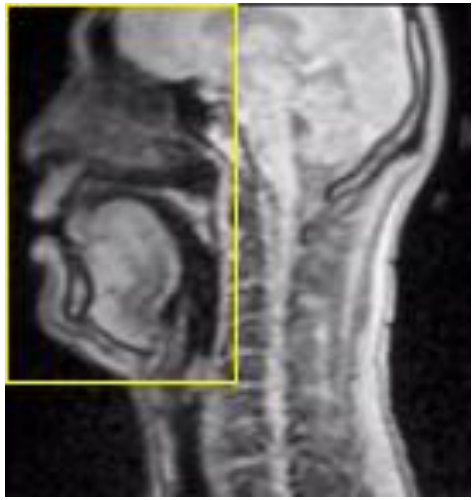
rtMRI	
画像サイズ	256 × 256 px
解像度	1 mm × 1 mm
スライス厚	10 mm
フレームレート	14 fps or 27 fps
音声	
標本化周波数	48 kHz
量子化ビット数	16 bit

音声データにおけるMRI撮像時のノイズは Zhaoら [3] の手法にて除去.

[3] Zhao et al., Proc. ICASSP 2022, 9281-9285 (2022).

データセット

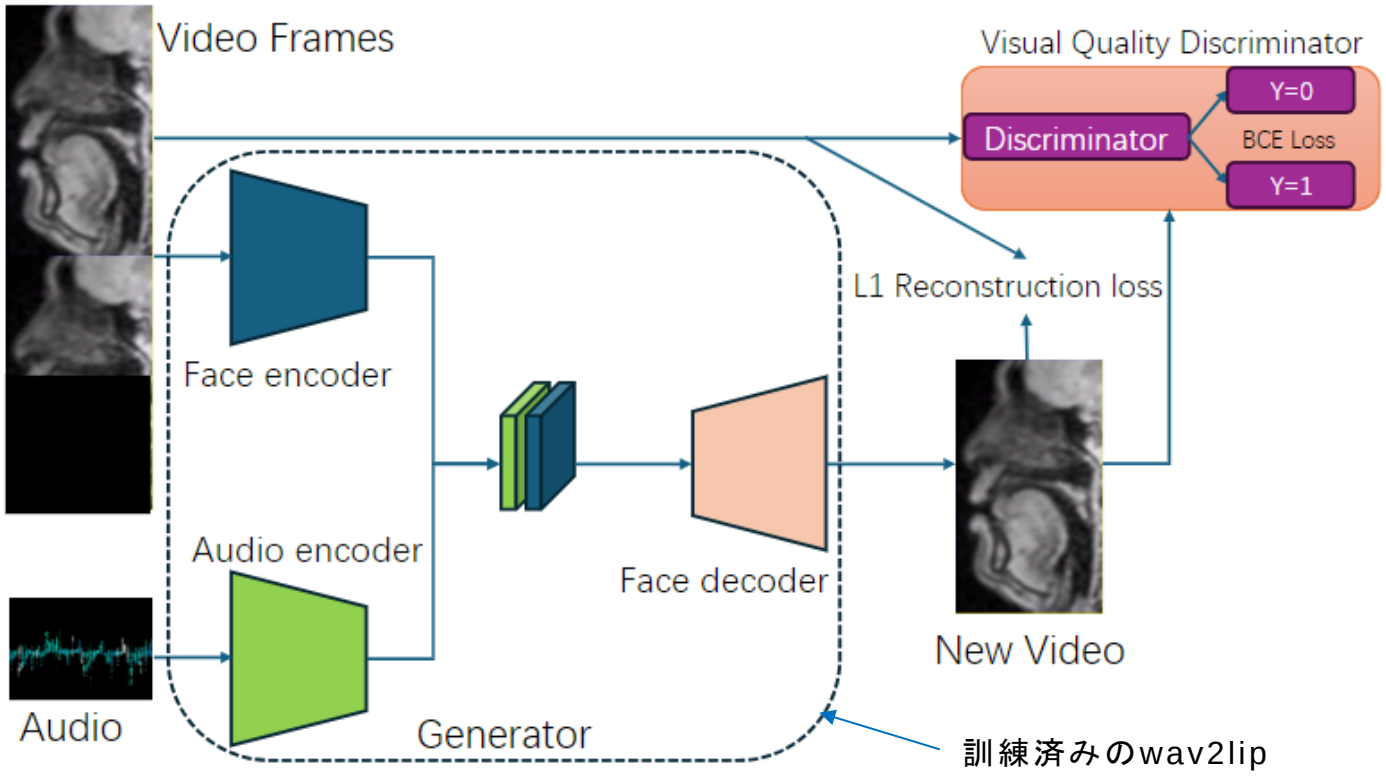
- ✓ オリジナル動画から**黄色**で囲んだ領域をトリミング
- ✓ フレー ムレートを25 fps, 画像サイズを1920 × 1080 pxに統一
- ✓ 動画を約5 sごとに分割



rtMRI image and trimmed area(yellow rectangle).

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

モデル学習の流れ



リップシンク動画生成モデル wav2lip [1] のネットワークと訓練済みのcheckpointを利用する.

Dataset のTrainset で, ファインチューニング.

Early Stopping を設置, L1 Reconstruction loss 値が 0.0025 付近で安定し訓練終了.

[1] Prajwal et al., Proc. 28th ACM International Conference on Multimedia, 484–492 (2020).

生成精度の評価(1/2)

- 3 組の検証データを準備した.
  条件 1 :Test セットからランダムに得たもの
  条件 2 :話者 W による「北風と太陽」日本語データ
  条件 3 :話者 W による「北風と太陽」中国語データ

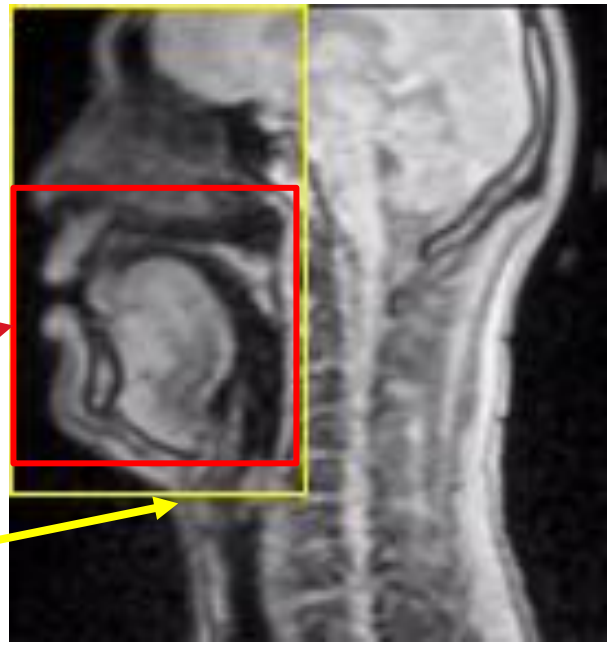
- 3 条件の検証データから各5つの動画を選んで，その音声と第1フレームの画像から，検証用動画を生成.
- Structural Similarity (SSIM) と Normalized Mean Squared Error (NMSE) で評価.

生成精度の評価(2/2)

生成動画とオリジナル動画の赤で囲んだ領域をトリミングし，SSIMとNMSEを計算.

評価領域400 x 420 px

学習と生成で使用した領域

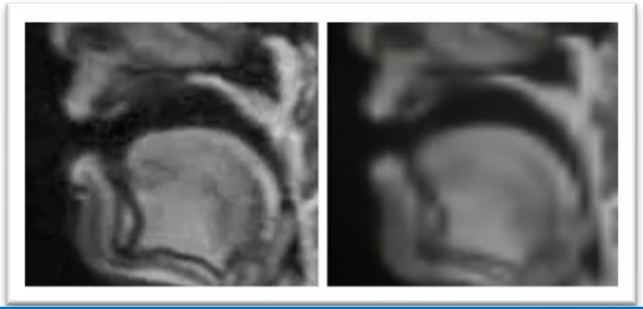


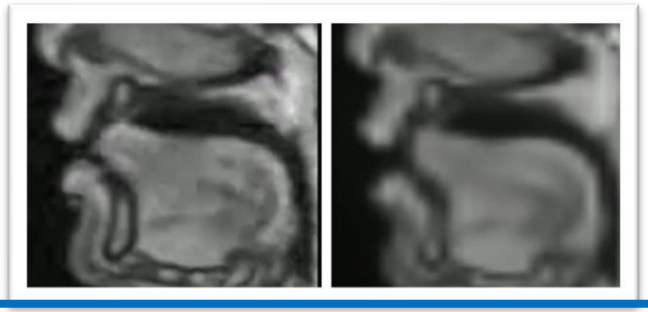
rtMRI trimmed area(yellow rectangle) and evaluation area (red rectangle).

結果

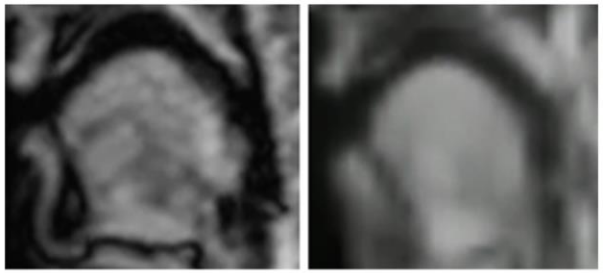
Original Generated Original Generated

Condition 1

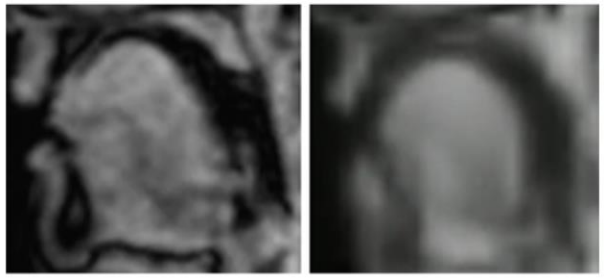




Condition 2



Condition 3



I. SSIMは1に近いほど評価が高い

II. NMSEは0に近いほど評価が高い

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

考察

① 提案法によってオリジナルに近い動画が生成された.

② モデルの学習に用いた話者の生成動画はSSIMとNMSEによる評価が高い.

③ 学習に用いなかった話者でも動作の生成は可能だが，画質は劣る．未学習の言語について，評価値の差が小さくなるため，生成精度の言語依存性が小さいと考えている．日本語には存在しない音について詳細に分析する必要がある.

おわりに

① 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 動画の生成を試みていい結果ができた.

② 未学習話者は学習済み話者と比べると生成動画の質が下がる，未学習言語は生成動画の質に影響が小さいという結論が生み出した.

③ 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，各言語に存在するユニークな発音は動画生成にどのような影響を与えることを分析する.

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.

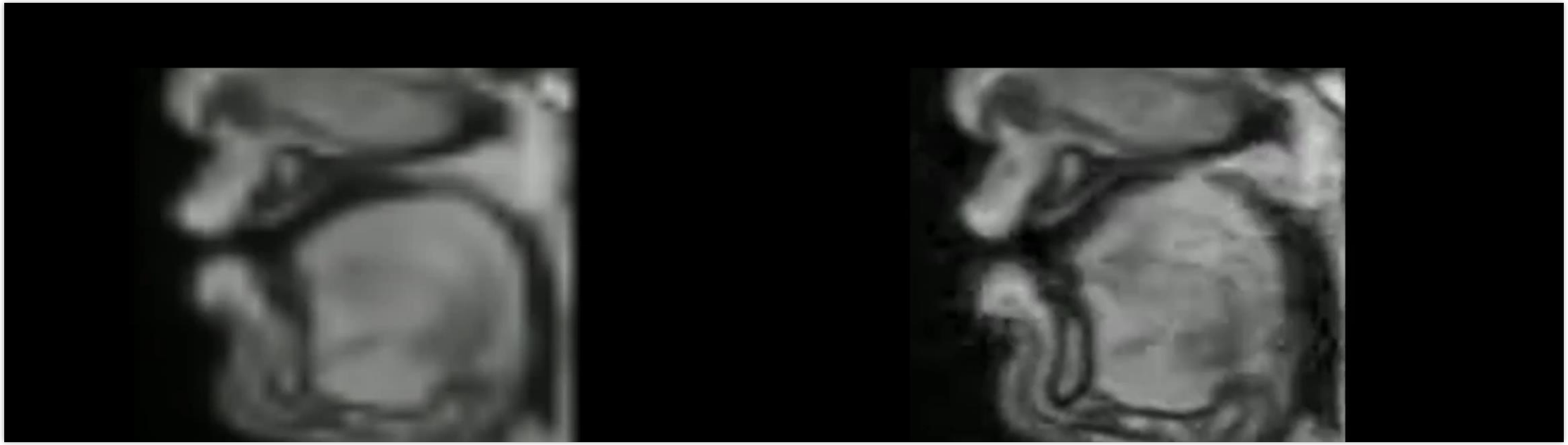
質疑応答用

結果

生成動画

オリジナル動画

結果



生成動画　　オリジナル動画

wav2lip

- L1Reconstruction loss.

$$L1\ Loss = \frac{1}{N}\sum_{i=1}^{N}|A_i - V_i|$$

- A_i：音声特徴
- V_i：口唇特徴

生成精度の評価

- Structural Similarity (SSIM) と Normalized Mean Squared Error (NMSE) で評価した.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

$$\mathrm{NMSE} = \frac{MSE(x_{true}, y_{perd})}{var(x_{true})}$$

生成精度の評価

$$\mathrm{MSE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(x[i][j] - y[i][j])^2$$