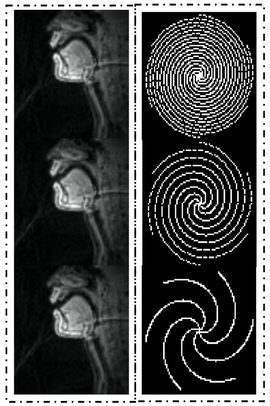
High Spatial-Temporal Resolution MRI Reconstruction for Speech Production Analysis

Stanford University

Qingyue Wei qywei@stanford.edu
Xuetong Zhou* xuetongz@stanford.edu

Problem/Background

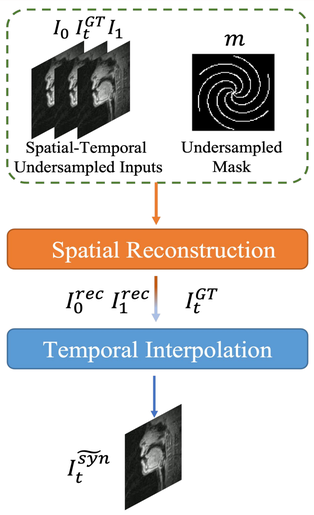
- Real-time magnetic resonance imaging (RT-MRI) is one of the modalities used to image the structures involved in speech production with the advantages of non-invasiveness and ability to images arbitrary planes and deep structures.
- Trade-off exists between spatial and temporal resolution.
- In this project, our goal is to reconstruct high spatial-temporal resolution MRI videos based on videos which are undersampled in both spatial and time domain.
- Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) are used for evaluation.



Datasets

- A multi-speaker dataset of speech production RT-MRI are used in this project. The dataset includes raw data and synchronized audio acquired during multiple speaking tasks on 75 speakers.
- Due to time & computation limitations, only includes one video for each speaker. Data is split into training/validation/test set (40K/10K/10K frames).
- Inputs are normalized ([0,1]) video frames undersampled in both spatial and time domain as well as the correspondning spatial undersampled mask.
- Direct outputs are the synthesized missing frames. And with all spatial and temporal reconstructed frames could produce high spatial and temporal resolution videos.

Method Overview

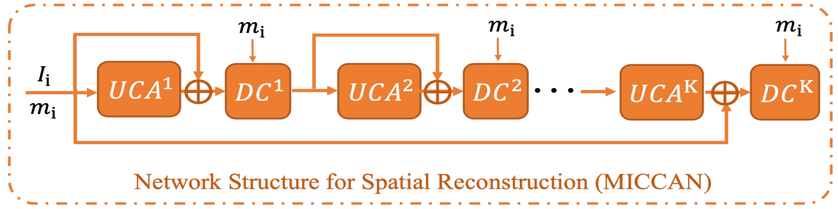


Our proposed pipeline is composed by two consecutive stages:

- Spatial Reconstruction
  - To restore high spatial resolution
  - Use an end-to-end model (MICCAN[1])
- Temporal Interpolation
  - To restore high temporal resolution
  - Use an end-to-end model (IFNet[2])

Model Details

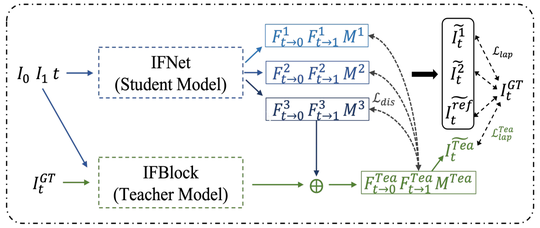
MICCAN for Spatial Reconstruction



Network Structure for Spatial Reconstruction (MICCAN)

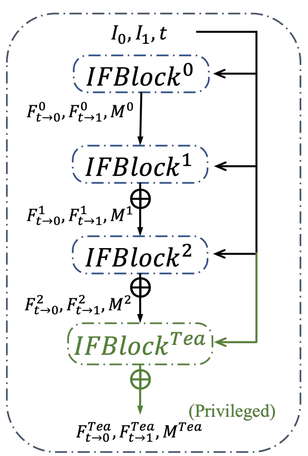
- MICCAN adopts several cascaded blocks to simulate the iterative process . Each block is composed by
  - UNet-based Channel-wise Attention (UCA) module and,
  - Data Consistency Layer (DC).
- The whole reconstruction process could be written as
  - $I_i^{rec} = F_K^S(F_{K-1}^S(\dots(F_1^S(I_i, m_i; \theta_1^S)\dots); \theta_{K-1}^S); \theta_K^S)$

IFNet for Temporal Interpolation



IFNet could

- Synthesize the intermediate frame from coarse to fine with less time cost.
- Also include another Teacher Model to provide more useful guidance for Student Model.



For Student Model

- Consists of three IFBlocks
  - From coarse to fine with different downsampling scales.
- The synthesis process could be written as
  - $\widetilde{I}_t^m = M^m \circ \widehat{I}_{t\to 0}^m + (1 - M^m) \circ \widehat{I}_{t\to 1}^m$
    $\widehat{I}_{t\leftarrow 0} = \overleftarrow{\mathcal{W}}(I_0, F_{t\to 0}), \quad \widehat{I}_{t\leftarrow 1} = \overleftarrow{\mathcal{W}}(I_1, F_{t\to 1})$

For Teacher Model

- Consists of one IFBlock.
- Distillation loss is adopted to provide more supervision that
  - $\mathcal{L}_{dis} = \sum_{m=1}^{3} \sum_{i\in 0,1} ||F_{t\to i}^{Tea} - F_{t\to i}^{m}||_1$

Results & Discussion

Method	Ratio	PSNR(dB)	SSIM
ADMM+TV	$arms6_1$	24.37	0.7830
MICCAN	$arms6_1$	**34.07**	**0.9157**
ADMM+TV	$arms6_2$	20.32	0.6023
MICCAN	$arms6_2$	**29.64**	**0.8330**
ADMM+TV	$arms6_6$	17.39	0.4910
MICCAN	$arms6_6$	**26.89**	**0.7630**

Table 1 spatial reconstruction results

Ratio	PSNR(rec)	SSIM(rec)	PSNR(gt)	SSIM(gt)
$arms6_1$	40.87	0.9938	30.22	0.9399
$arms6_2$	41.41	0.9949	26.77	0.8846
$arms6_6$	40.86	0.9940	24.07	0.8326
original	-	-	39.33	0.9928

Table 2 spatial & temporal reconstruction results

Experiments show that

- MICCAN outperforms ADMM+TV by a large margin (~10dB) under different undersampled ratios.
- MICCAN is able to deal with severe spatial undersampled inputs ($26.89_{(MICCAN)}$ vs. $17.39_{(ADMM+TV)}$ in PSNR).
- Our spatial & temporal reconstruction pipeline is robust to different undersampled ratios and shows promising results.
- For temporal interpolation, our model could accurately predict the movements of the tongue while speaking.

Take away

- Performances of temporal interpolation results rely heavily on the previous spatial reconstructed results ⇒ could spend more time for acquiring MR images to get higher spatial resolution frames.
- Using auxiliary loss (e.g., distillation loss, perceptual loss) could help improve model performance.

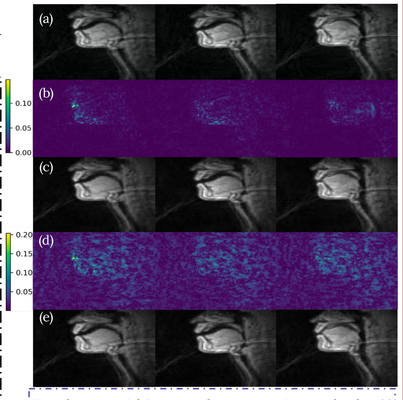


Fig 1 shows spatial & temporal reconstruction results that (a) spatial reconstructed GT; (c) spatial & temporal reconstructed outputs; (e) GT from raw data; (b) error maps between a&c; (d) error maps between e&c

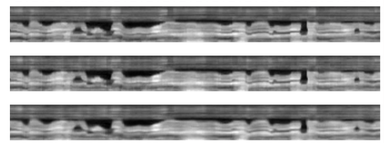


Fig 2 shows tongue movements in one speaking video that the first/second/third row indicates tongue movements in the raw/spatial reconstructed/spatial & temporal reconstructed video

Conclusion & Future work

- Our proposed pipeline provide a way to reconstruct a high spatial and temporal resolution MRI videos. Moreover, we found out that spatial resolution of the undersampled input data plays an important role for the final performance of our interpolation network. Therefore, if we spend more time for MRI scanning and downsampled more along time dimension, we could get higher spatial resolution frames and IFNet could then generate videos with better quality (i.e., higher spatial & temporal resolution).
- Future work could include
  - Conducting more experiments on different datasets,
  - Seeking for more powerful spatial reconstruction networks,
  - Trying to increase the interval between two frames (increasing downsampling ratio in time domain) and synthesize multiple intermediate frames and so on.

Reference

[1] Huang, Qiaoying, et al. "MRI reconstruction via cascaded channel-wise attention network." *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019.
[2] Huang, Zhewei, et al. "Rife: Real-time intermediate flow estimation for video frame interpolation." *arXiv preprint arXiv:2011.06294* (2020).