THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Deep Learning for Ultrasound-Based Tongue Contour Segmentation and Speech Disorder Classification

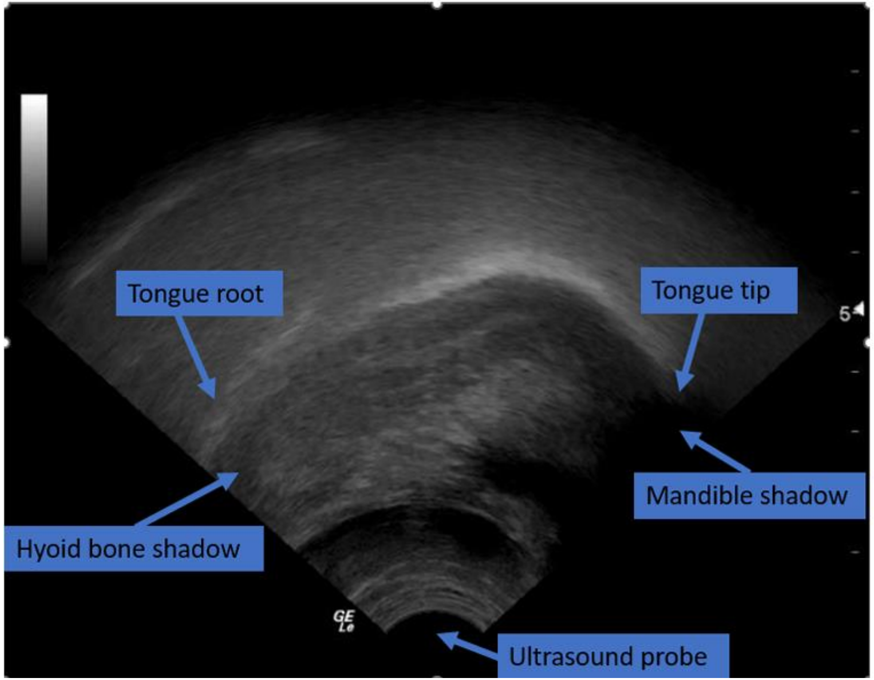
Alisher Myrgyyassov[1], Stefanie Sun[1,2], Zhen Song[1], Bruce Wang[3], Min Ney Wong[4], Yongping Zheng[1,2,*]

[1]Department of Biomedical Engineering, [2]Research Institute for Smart Ageing, [3]Department of English and Communication, [4]Department Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

Introduction

Tongue Contour Significance

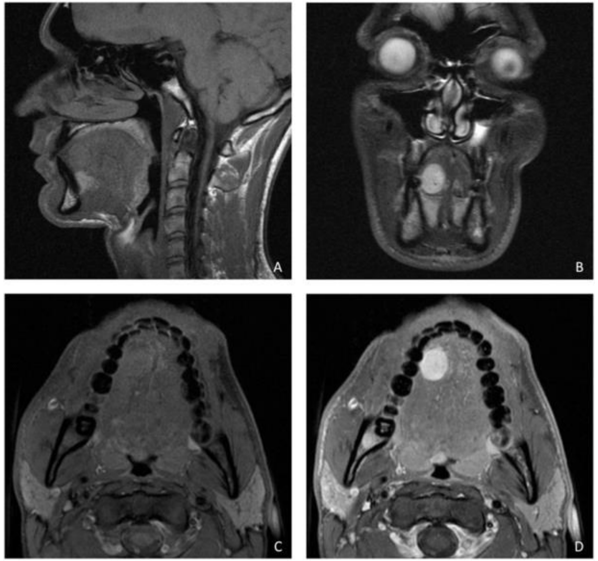


Ultrasound Image of the tongue
(Al-hammuri et al., 2022)

Tracking the tongue contour in biomedical imaging provides essential information about the **kinematics and shape** of the tongue during speech.
(Karimi et al., 2019)

This kinematic data can potentially be used for speech assessment and speech disorder classification.

Modality Choice (1)



MRI Tongue Imaging example
(Abreu et al., 2017)

Pros:

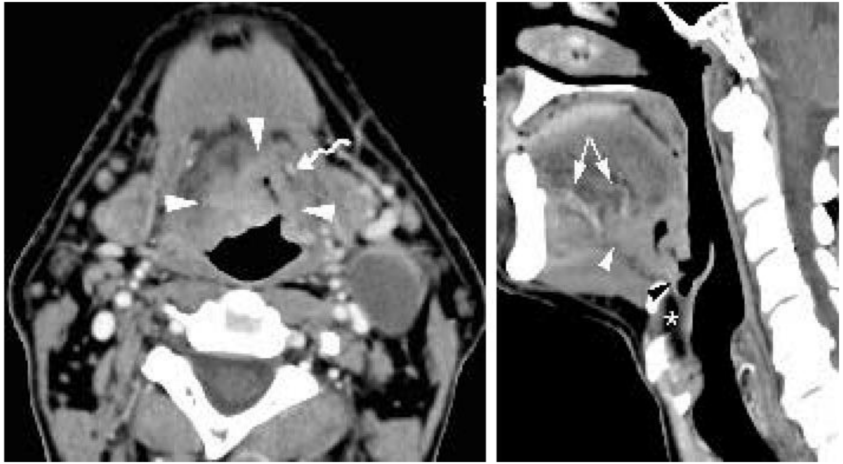
- Real-time acquisition
- High resolution 3D image
- High contrast between soft tissues

Cons:

- Expensive
- Large-sized
- Long acquisition time

Therefore, MRI is not suitable for clinical studies of speech disorders using imaging data.
(Al-hammuri et al., 2022)

Modality Choice (2)



CT Scan Tongue Imaging example
(van den Brekel & Castelijns, 2005)

Pros:

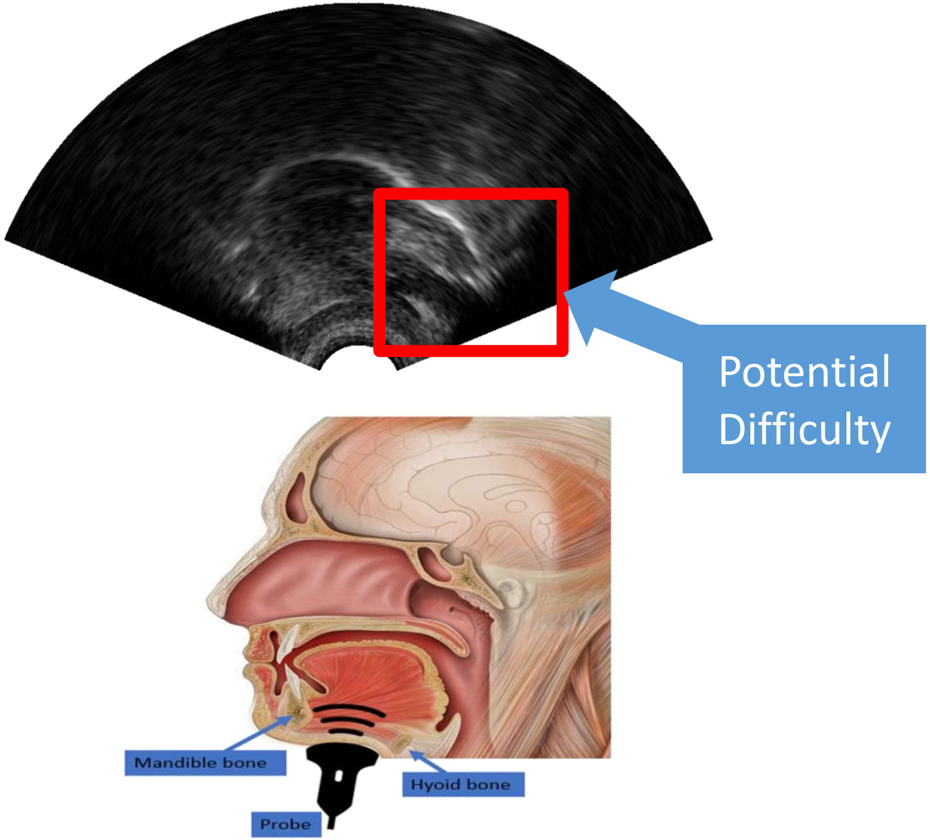
- Relatively cheap imaging solution
- Reasonable 3D imaging resolution
- High contrast between soft tissues

Cons:

- **Radiation danger**

CT and X-Ray **are widely used** in advanced surgical procedures related to vocal tract, however, **not suitable** for real-time day-to-day speech analysis. (Al-hammuri et al., 2022)

Modality Choice (3)



Ultrasound Imaging:

- Safe (no ionizing radiation)
- Rapid real-time data acquisition
- Cheap

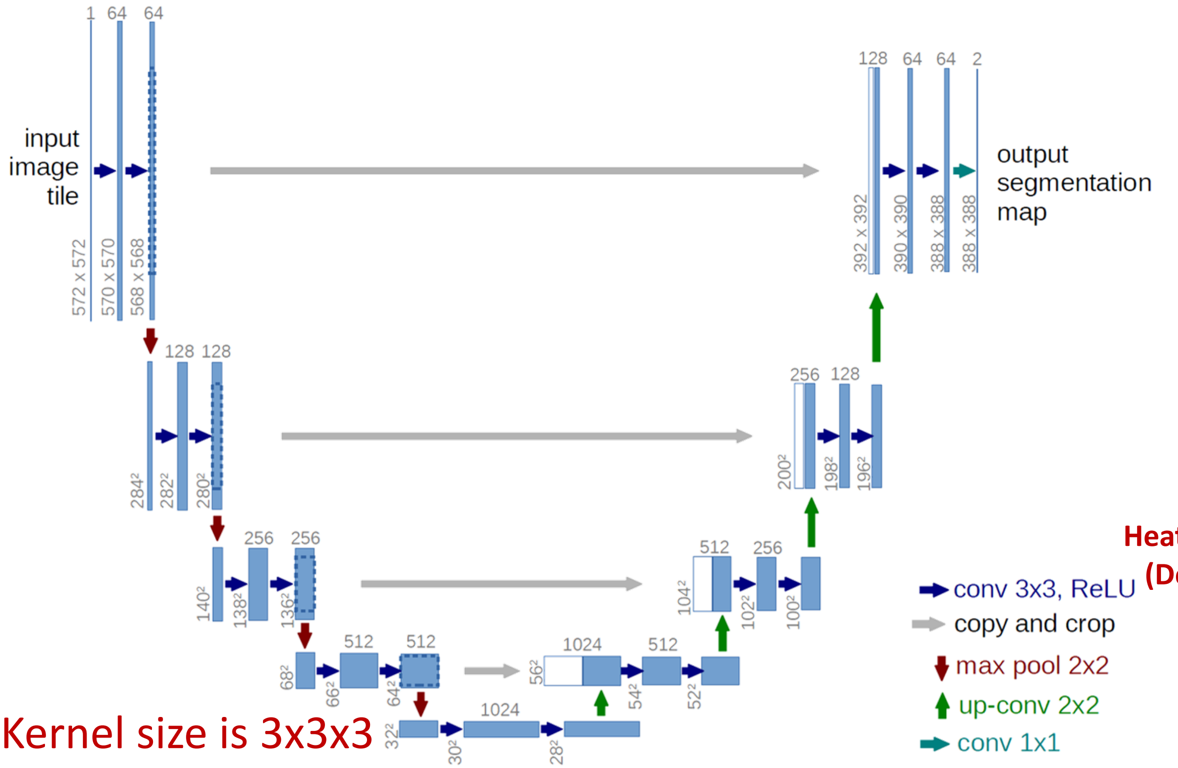
Therefore, using ultrasound imaging is considered to be the most safe and efficient method for speech assessment.

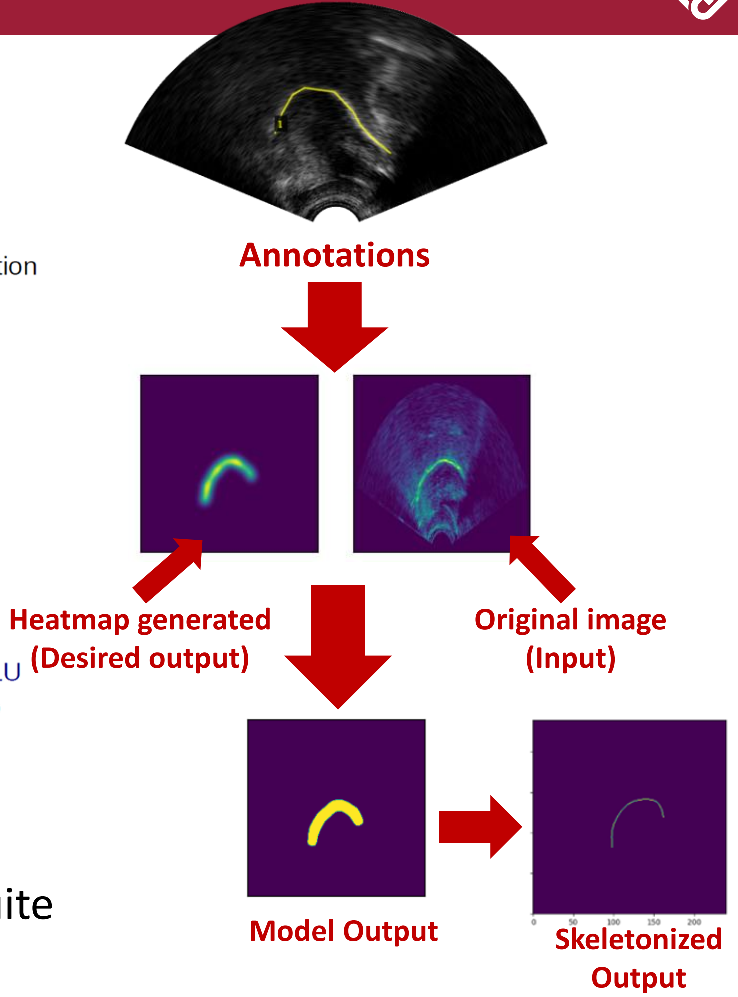
However, the imaging modality tends to have a high level of **ultrasound artifacts presence**.

(Al-hammuri et al., 2022)

Methodology

Deep Learning Model Selection





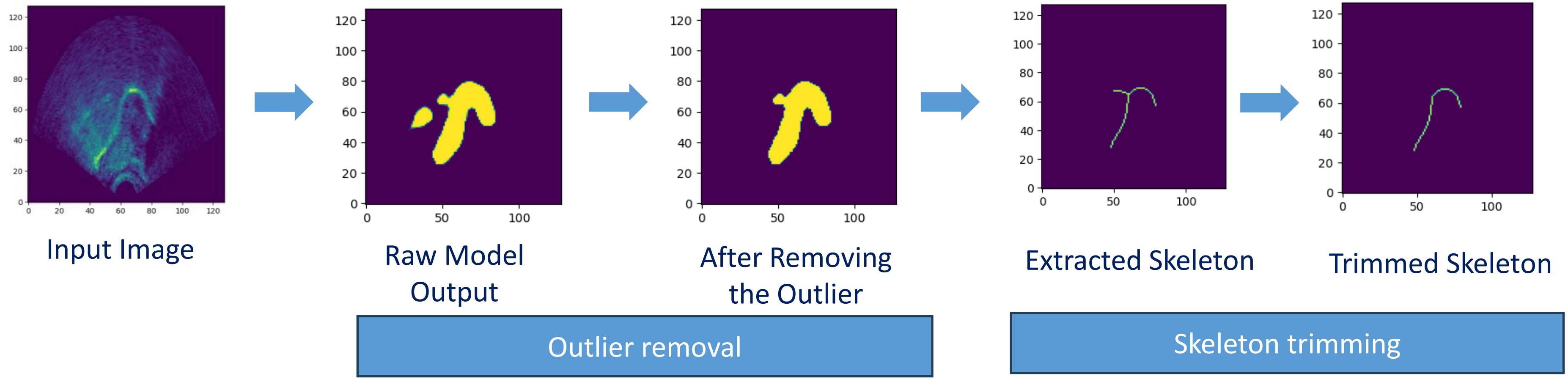
Kernel size is 3x3x3

More than 2000 images were annotated from the Ultrasuite Dataset by Eshky et al. (2018).

Training and Test Pipeline

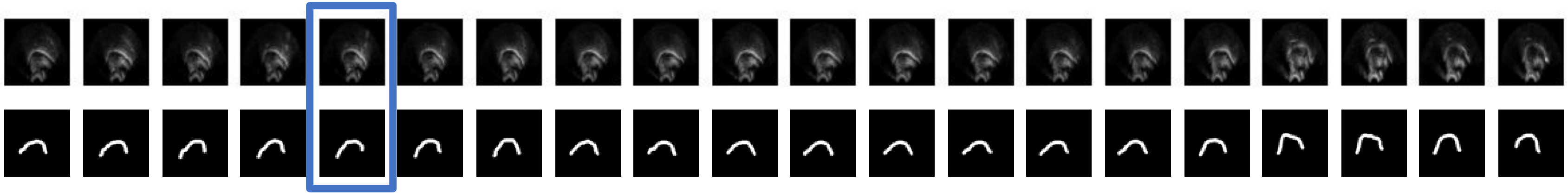
Summary of the key points:

- The full dataset was split into train, test, and validation datasets with the ratio of 80%, 10%, and 10%, respectively
- Test dataset is the previously unseen data
- We use different post-processing techniques to make the models output more natural, such as the outlier removal and skeleton trimming.

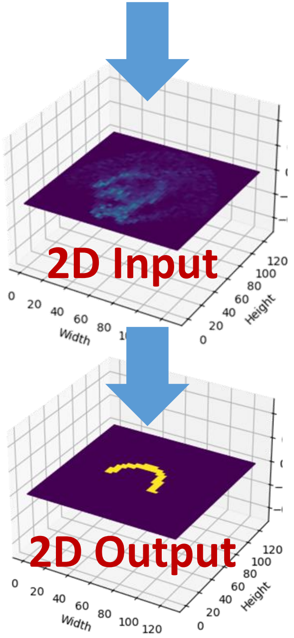


Results and Discussion

Input Selection (1)



Select one



This is the traditional approach commonly used in the ultrasound imaging segmentation field.

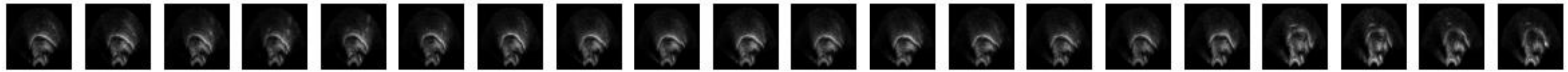
For example, Zhu et al. (2019) or Karimi et al. (2019)

Then, the model performance is assessed using the Mean Sum Distance (MSD) score:

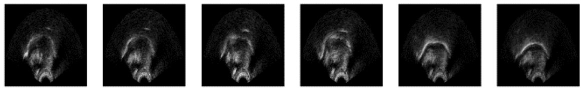
$$D(U,V) = \frac{1}{2n}\left(\sum_{i=1}^{n} \min_{j} |v_i - u_j| + \sum_{i=1}^{n} \min_{j} |u_i - v_j|\right)$$

MSD score is given in pixels and signifies the difference between the actual and desired outputs.

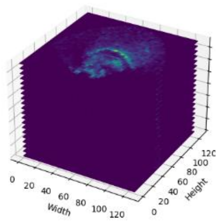
Input Selection (2)



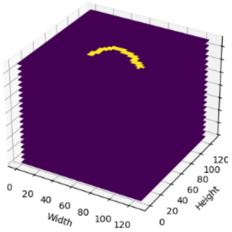
Fixed Length (N)



3D Input



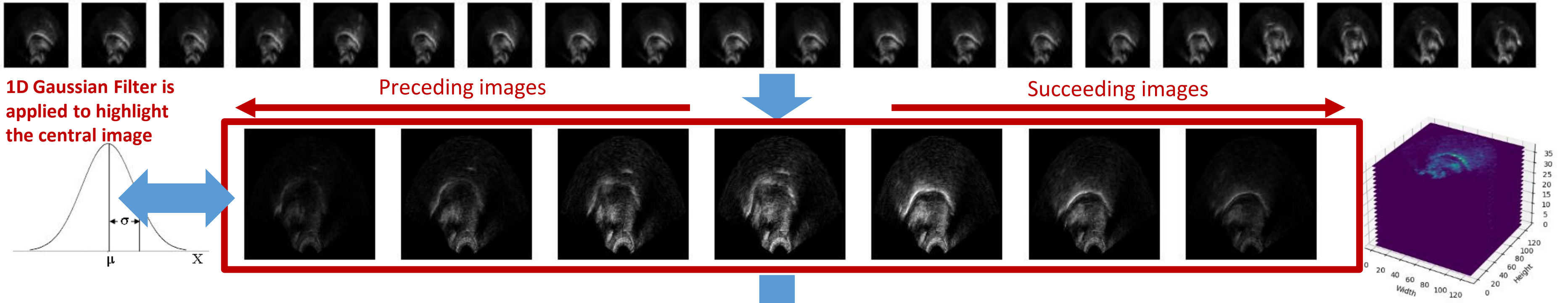
3D Output



Another common approach is to stack multiple 2D images into a single **N-stack** 3D tensor, where N is the number of 2D frames stacked together.

However, the first and the last images **do not have contextual information** about preceding and succeeding images, respectively.

Proposed Methodology



The multichannel approach results:

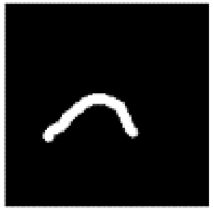
- 1.37px on different speakers' data resulting in **1.28mm** with about 2000 annotated frames.

In comparison to:

- **1.43mm** achieved by Zhu et al. (2019) on the same dataset with 17580 annotated frames.

2D 7-Channel Input

2D Output

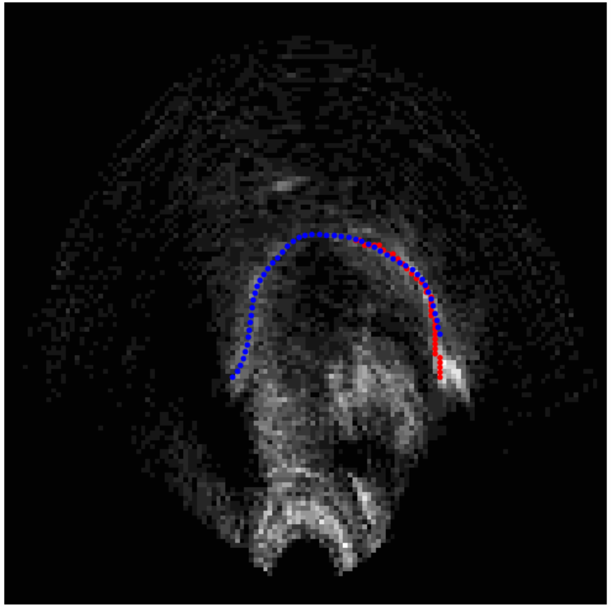


- 1.39px before post-processing
- **1.37px** after post-processing

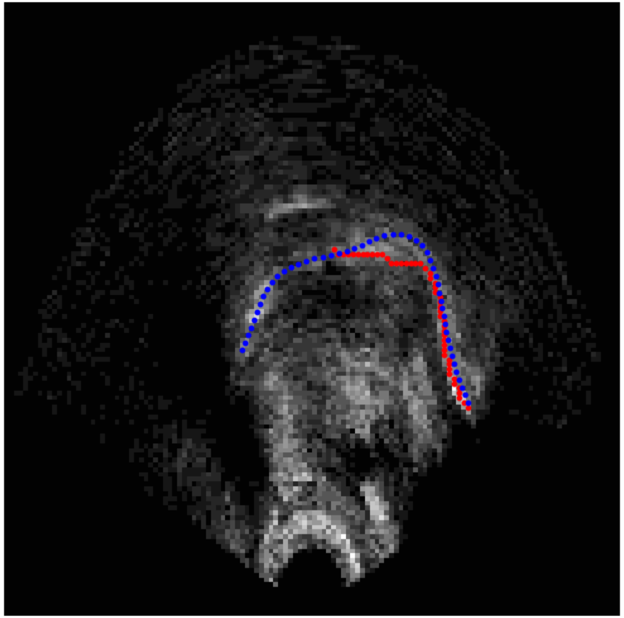
The post-processing also does not improve the performance of the model significantly, meaning that the model gives more stable and natural results

Results

The animations demonstrate a visual comparison between two approaches.



2D Approach



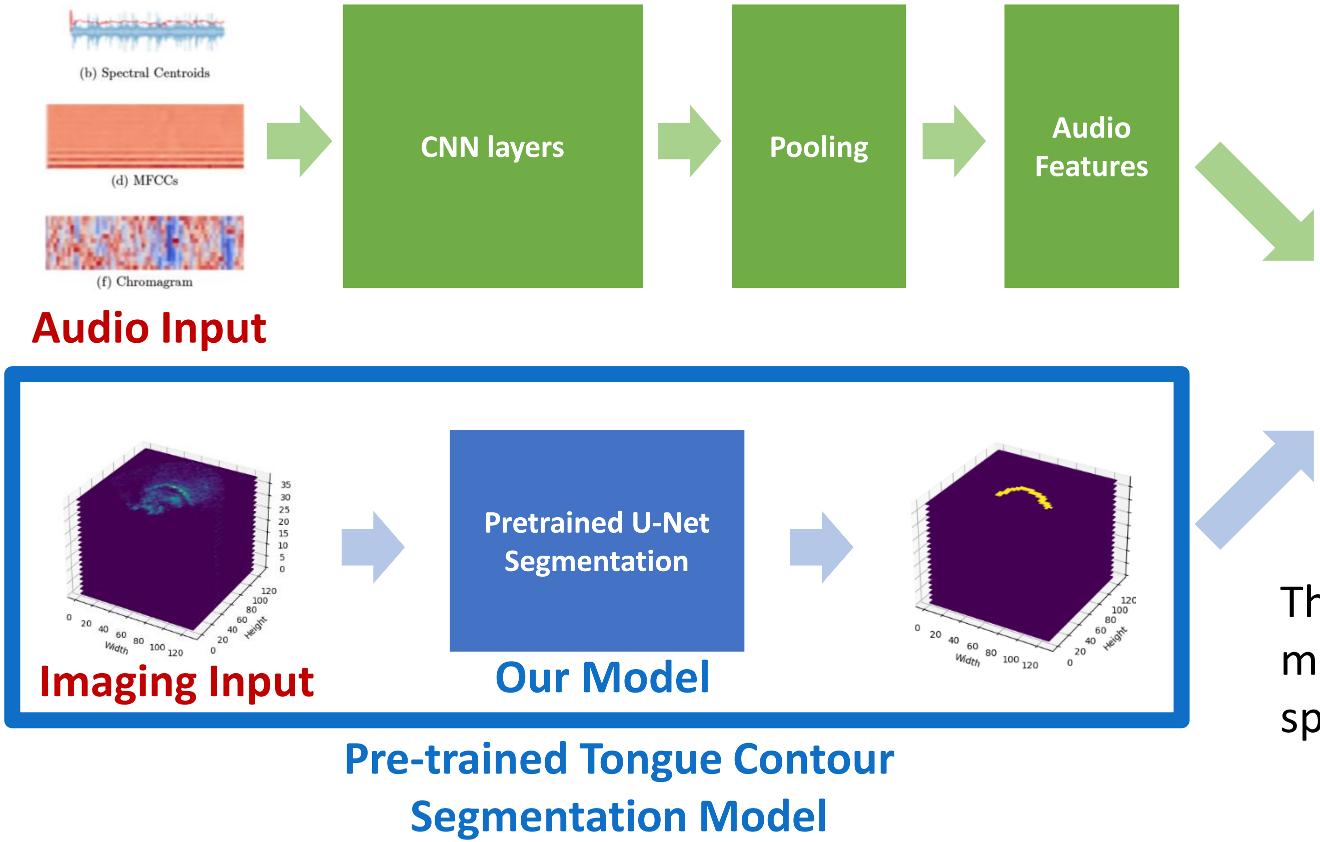
Multichannel Approach

Results Summary For Same Speakers

	No post-processing (px)	Post-processing (px)	Dice Score
2D	2.44	1.49	0.83
9-stack	1.77	1.68	0.83
12-stack	1.91	1.55	**0.84**
7-channel	**1.39**	**1.38**	0.83

Conclusion and Future Work

Fully-Automated Speech Disorder Classification



The proposed model uses multimodal data to classify speech disorders.

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Thank You!

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

Appendix

	No post-processing	Trim	Largest	Both	Dice Score
9-stack same-speaker	1.32	1.42	1.46	1.45	0.86
9-stack different-speaker	1.77	1.75	1.68	1.68	0.83
12-stack same-speaker	1.72	1.67	1.51	1.50	0.85
12-stack different-speaker	1.91	1.54	1.55	1.55	0.84
2D same-speaker	1.39	1.23	1.24	1.24	0.87
2D different-speaker	2.44	1.74	1.51	1.49	0.83
7-channel same-speaker	1.38	1.44	1.49	1.48	0.87
7-channel different-speaker	**1.37**	**1.37**	**1.37**	**1.37**	**0.83**
5-channel different-speaker	1.42	1.42	1.41	1.41	0.82
5-channel same-speaker	1.42	1.47	1.47	1.47	0.87

References

- Abreu, I., Roriz, D., Rodrigues, P., Moreira, A., Marques, C., & Caseiro-Alves, F. (2017). Schwannoma of the tongue—A common tumor in a rare location: A case report. European Journal of Radiology Open, 4, 1-3. https://doi.org/10.1016/j.ejro.2017.01.002
- Al-hammuri, K., Gebali, F., Thirumarai Chelvan, I., & Kanan, A. (2022). Tongue Contour Tracking and Segmentation in Lingual Ultrasound for Speech Recognition: A Review. Diagnostics, 12(11), 2811. https://doi.org/10.3390/diagnostics12112811
- Eshky, A., Ribeiro, M., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J., & Wrench, A. (2018). UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions. In Proceedings of Interspeech (pp. 1888-1892). https://doi.org/10.21437/Interspeech.2018-1736
- Karimi, E., Ménard, L., & Laporte, C. (2019). Fully-automated tongue detection in ultrasound images. Computers in Biology and Medicine, 111, 103335.
- van den Brekel, M., & Castelijns, J. (2005). What the clinician wants to know: surgical perspective and ultrasound for lymph node imaging of the neck. Cancer Imaging: The Official Publication of the International Cancer Imaging Society, 5(Spec No A), S41-S49. https://doi.org/10.1102/1470-7330.2005.0028
- Zhu, J., Styler, W., & Calloway, I. (2019). A CNN-based tool for automatic tongue contour tracking in ultrasound images. Department of Linguistics, University of Michigan, United States.