Refining rtMRI Landmark-Based Vocal Tract Contour Labels with FCN-Based Smoothing and Point-to-Curve Projection

Mushaffa Rasyid Ridha[1] and Sakriani Sakti[1,2]

[1]Japan Advanced Institute of Science and Technology (JAIST), Japan
[2]Nara Institute of Science and Technology (NAIST), Japan

Introduction

Motivation Background

- **RtMRI of Articulatory Movement**
  - → Advanced real-time magnetic resonance imaging (rtMRI) allows researchers to study articulatory movements during speech production with high temporal resolution
  - → For in-depth analysis of dynamic movements, **detecting articulatory contour is essential**
  - → **Problem:** Manual labeling of articulatory contour is nearly impossible

- **Existing USC-TIMIT rtMRI dataset** (Bresch and Narayanan,2009)
  - → Provides rtMRI data with landmark-based contour labels for part of the data
  - → Labels were derived from **unsupervised region segmentation** using spatial frequency domain representation and gradient descent optimization
  - → **Problem:**
    - ✓ While this method yields high-quality labels, occasional labeling errors exist
    - ✓ Many contour detection methods were trained based on this ground truth, which is not purely a gold standard label

Refinement of the label of rtMRI data is critical

Contribution of this paper

- **Refinement of the landmark-based vocal tract contour labels, with steps:**

  → Outlier removal

  → The utilization of a fully convolutional network (FCN) to smooth contour shapes

  → A point-to-curve projection technique to fit the edge plane of the articulators

- **Investigation of the quality of the new labels through subjective assessments**

  → Comparing them to the existing data labels and a comprehensive analysis regarding which contour labels are most prone to issues

Related Works in Estimating Vocal Track Contour

- **Landmark-based Approaches**
  - → Active Control Model (ACM) (Kass et al., 1988)
  - → Active Shape Model (ASM) (Silva and Teixeira, 2015)
  - → Articulatory-Specific Multiple Linear Regression (MLR) (Labrunie et al., 2018)
  - → Unsupervised Region Segmentation with Spatial Frequency Domain Representation and Gradient Descent Optimization (Bresch and Narayanan, 2009)
- **Segmentation-based Approaches**
  - → Many methods for Vocal Tract Segmentation (Raeesy et al., 2013; Ruthven et al., 2021) rely on supervised deep learning
  - → Multi-task learning for contour detection and labeling (Silva and Teixeira, 2015)

While many studies have introduced novel techniques of vocal track contour estimation, they often rely on existing ground truth, which may not be perfect.
In contrast, this paper enhances the ground truth to contribute to the community.

Database Description

Dataset

USC-TIMIT (Bresch and Narayanan,2009)

- Synchronized audio and rtMRI data
- 10 natives speakers of General American English
- 460-sentence phonetically balanced dataset used in the MOCHA-TIMIT corpus (Wrench, 1999)
- The visual articulatory MRI data with frame rate of 23.18 frames per second

Dataset

USC-TIMIT (Bresch and Narayanan,2009)

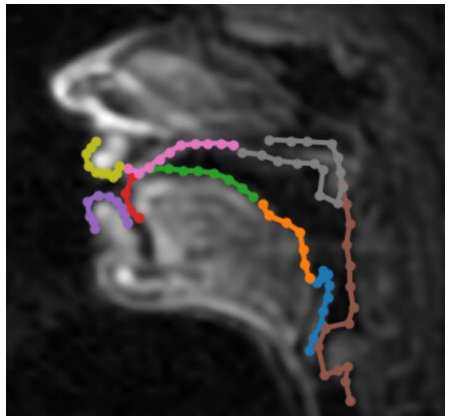
- Synchronized audio and rtMRI data
- 10 natives speakers of General American English
- 460-sentence phonetically balanced dataset used in the MOCHA-TIMIT corpus (Wrench, 1999)
- The visual articulatory MRI data with frame rate of 23.18 frames per second

Only part of the data has the available landmark label

Dataset

USC-TIMIT (Bresch and Narayanan,2009)

- Synchronized audio and rtMRI data
- 10 natives speakers of General American English
- 460-sentence phonetically balanced dataset used in the MOCHA-TIMIT corpus (Wrench, 1999)
- The visual articulatory MRI data with frame rate of 23.18 frames per second
- There are 95,223 video frames in total



= Different format points, different segmentation

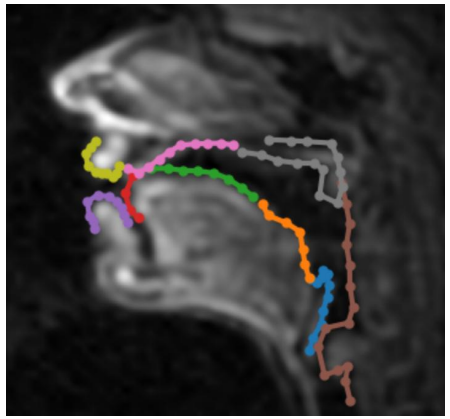
Format Points	# Frames (total)	# Frames (each subject)
178	15205	15205 (F1)
180	51252	16882 (F1) 34370 (F5)
181	28766	28766 (M3)

These labels are generated using spatial frequency domain-based segmentation and are available in different formats

Dataset

USC-TIMIT (Bresch and Narayanan,2009)

- Synchronized audio and rtMRI data
- 10 natives speakers of General American English
- 460-sentence phonetically balanced dataset used in the MOCHA-TIMIT corpus (Wrench, 1999)
- The visual articulatory MRI data with frame rate of 23.18 frames per second
- There are 95,223 video frames in total



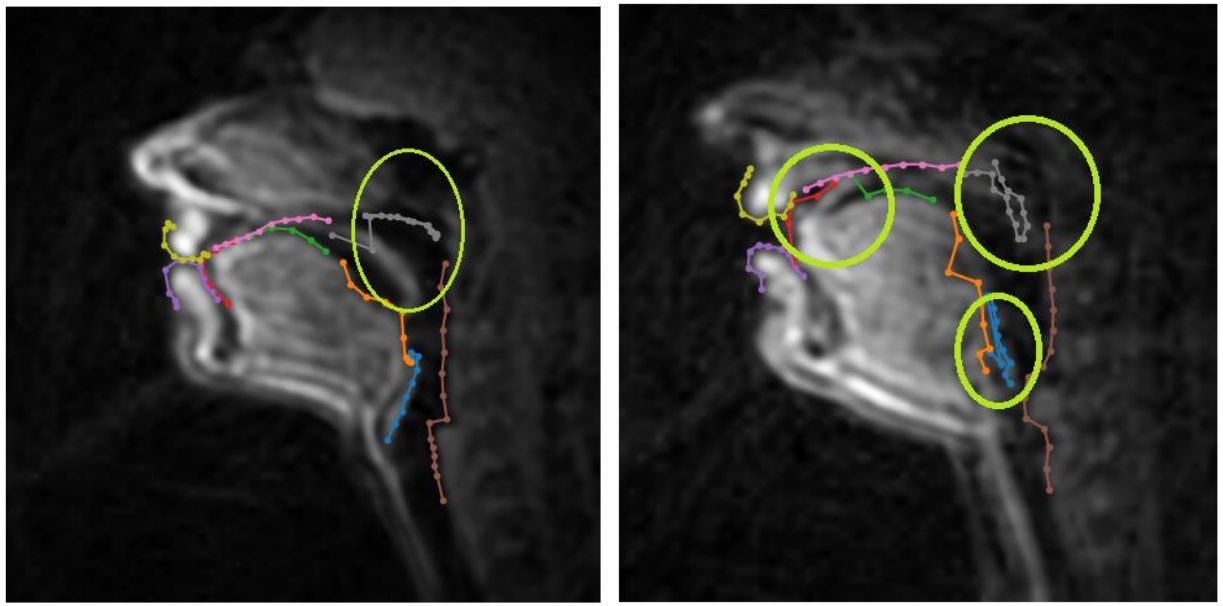
= Different format points, different segmentation

Format Points	# Frames (total)	# Frames (each subject)
178	15205	15205 (F1)
180	51252	16882 (F1) 34370 (F5)
181	28766	28766 (M3)

We primarily use the 180-point data, which constitutes over 50% of the available data

Proposed Label Refinement

Step 1: Outlier Removal



- **Objective:**
  → Address data outliers

- **Solution:**
  → Calculate the average size of landmark's area (ex. uvular's area) for every area
  → Center the data with the average size for each area
  → Remove data if too small or too big using a constant threshold value

We detected and removed 248 outliers from the F1 dataset and 2,757 outliers from the F5 dataset

Step 2: FCN-based Smoothing

- **Objective:**

  → Address noise-sensitive contour shapes

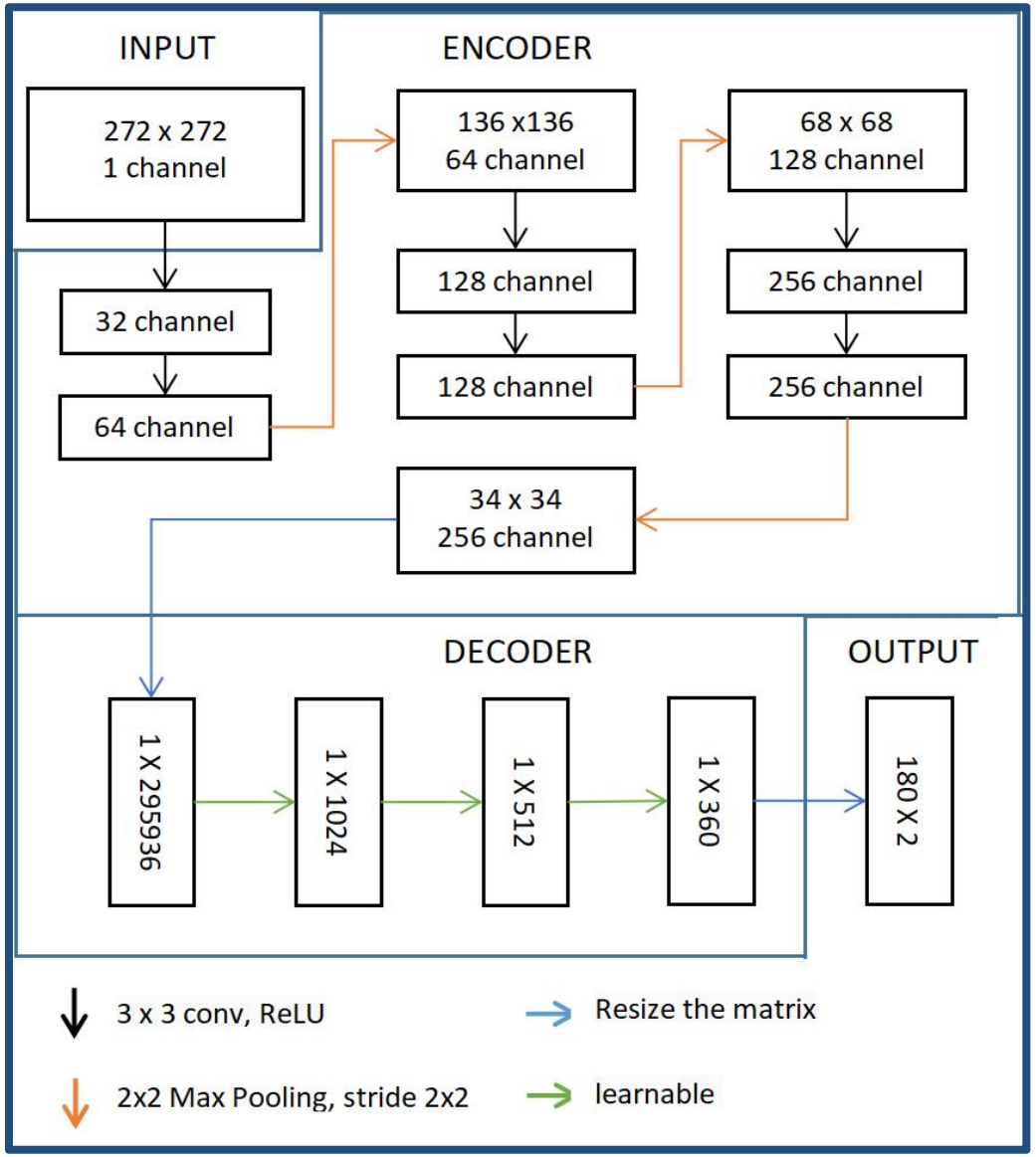
  → Given the rtMRI image data, produce smoother versions of the coordinate points for 180 landmarks

- **Solution:**

  → Neural networks are effective at handling labeling errors

  → Simpler models exhibit greater resilience to input noise, as complex models are more prone to overfitting and sensitivity to noisy inputs

  → Inspired by the U-Net approach (Ronneberger et al., 2015), we have implemented a more straightforward version of U-Net-based FCN



Step 3: Point-to-Curve Projection

- **Objective:**

  → Address noise-sensitive contour shapes due to the blurry images

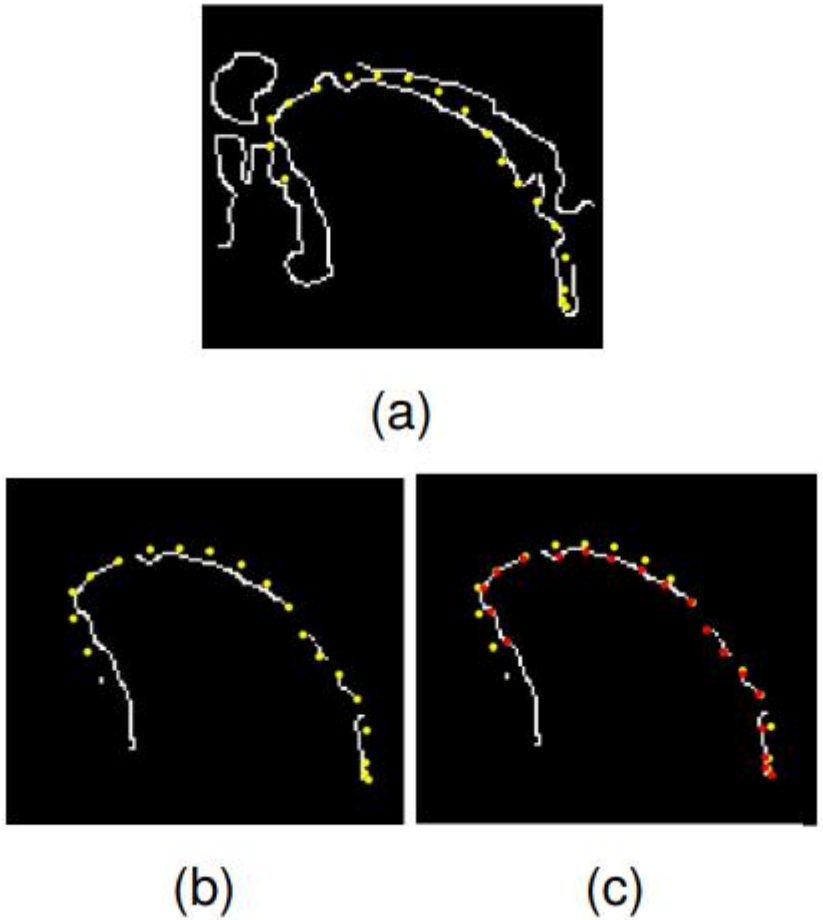
  → Project landmark points onto the edge plane of the articulators

- **Solution:**

  (a) Generate the edge of the MRI image using the adaptive threshold Gaussian method (Gaur et al., 2014)

  (b) Eliminate unnecessary edges that are too far from the smoothed points generated by FCN

  (c) Project the FCN points onto the nearest neighbor edge pixels to fit them onto the edge curve of the articulator



Evaluation

Experimental Set-Up

- **Subjective Evaluation**

→ A subjective evaluation involving 20 participant

→ This evaluation entailed an A/B preference test

→ Three different label sets generated from different methods:

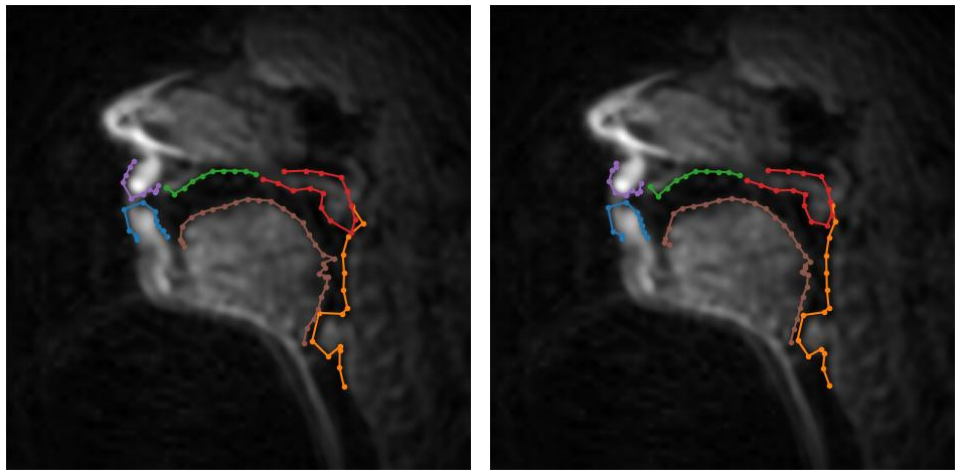
1. The original ground truth

2. FCN smoothing

3. FCN smoothing + Point-to-edge projection

→ 60 question randomly selected (3 sets of 20 questions)

To compare the quality of two randomly selected systems with two images (A and B)

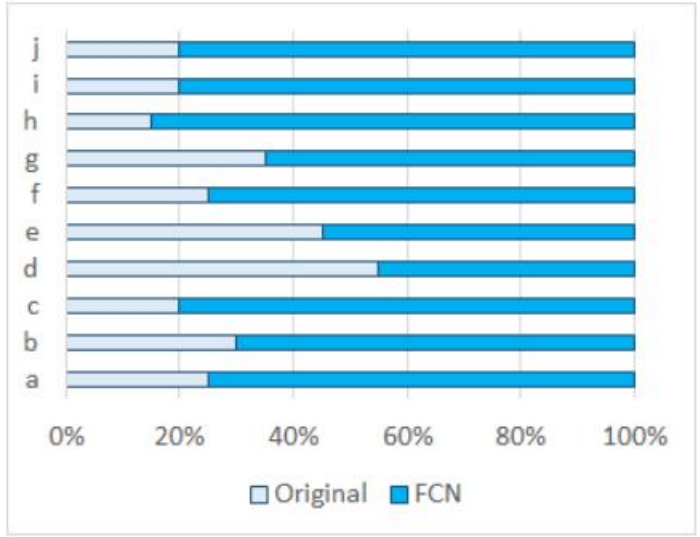


	A	B
a. Upper lip (purple)		(A/B)
b. Bottom lip (blue)		(A/B)
c. Palate / Hard Palate (Green)		(A/B)
d. Edge / Tip Tongue (Brown, left)		(A/B)
e. Middle Tongue (Brown, middle)		(A/B)
f. Back Tongue (Brown, right top)		(A/B)
g. Epiglottis (Brown, right bottom)		(A/B)
h. Uvular (Red)		(A/B)
i. Pharyngeal Wall (Orange)		(A/B)
j. Overall landmark		(A/B)

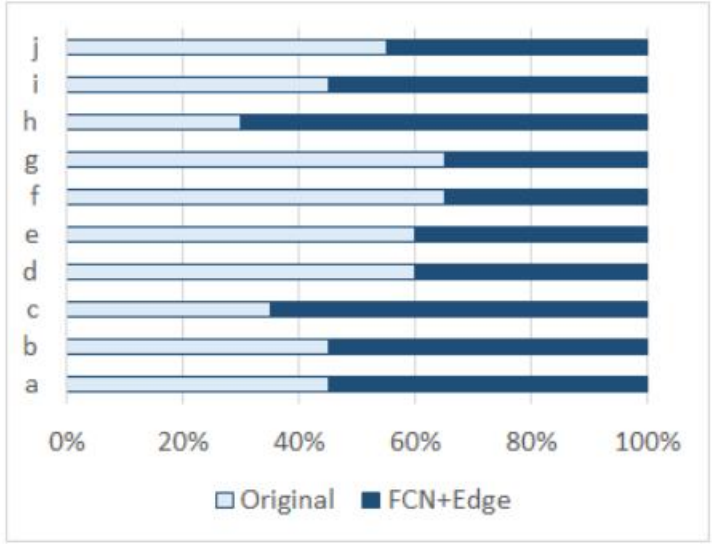
Experimental Results

AB Preference Tests

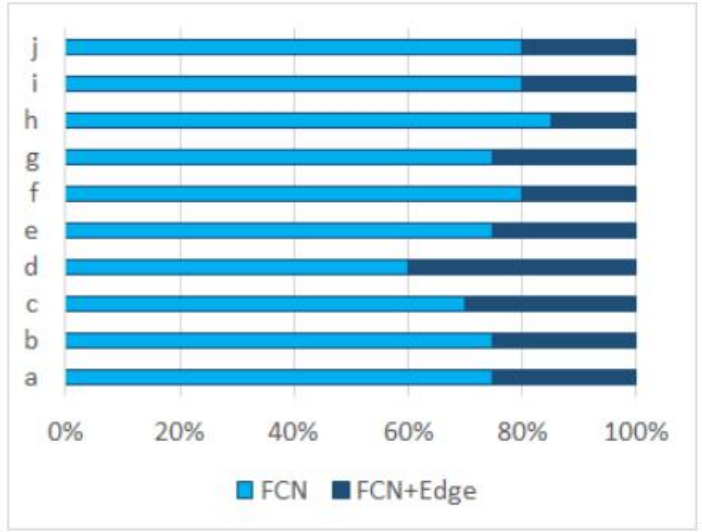
Considering nine local areas: (a) upper lip, (b) bottom lip, (c) hard palate, (d) edge of the tongue, (e) middle tongue, (f) back of the tongue, (g) epiglottis, (h) uvular, (i) pharyngeal wall and (j) the overall landmark-based vocal tract



(a) Original vs. FCN



(b) Original vs. FCN+Edge

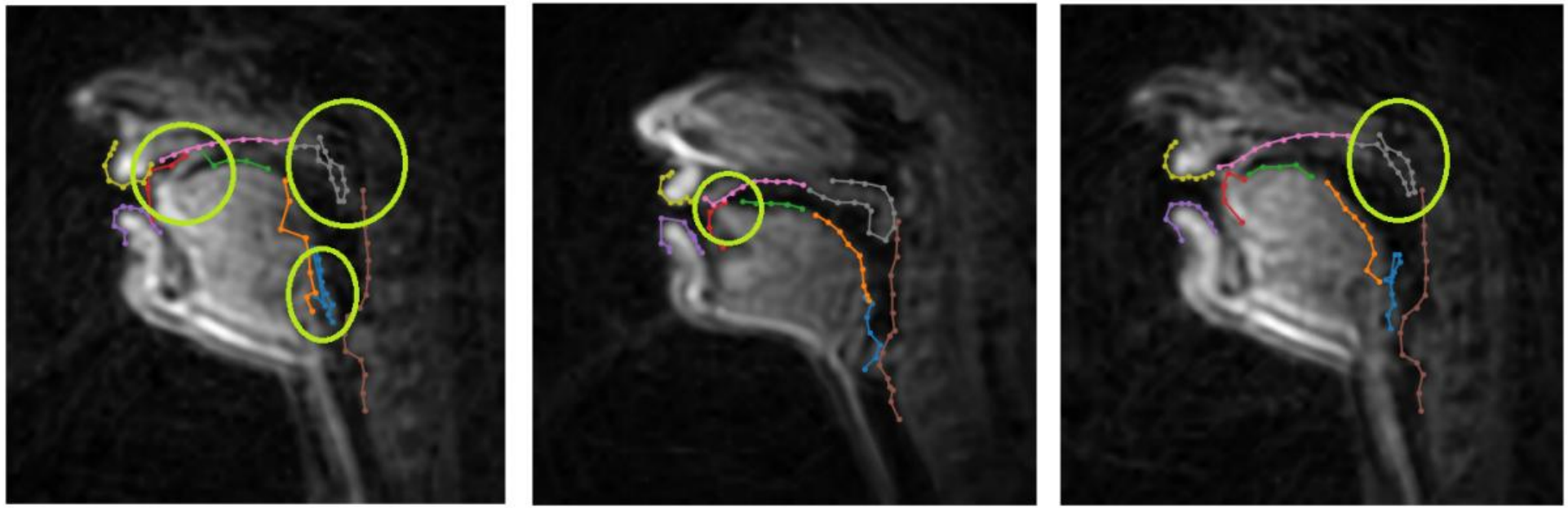


(c) FCN vs. FCN+Edge

The subjects judged that the proposed FCN labels are better than the original labels and FCN+Edge label data in almost all areas

Experimental Results

- **Errors and inaccuracies of the landmark contour labels**



(a) Original (b) FCN (c) FCN+Edge

→ The most critical area prone to errors in the original data is the uvular area
→ The new labels, enhance accuracy and reliability

Conclusion

Conclusion

- This paper contributes to the field by offering a refinement of landmark-based vocal-tract contour labels of rtMRI USC-TIMIT dataset

- This refinement includes:
  → **Outlier removal, FCN-based smoothing, and a landmark point-to-edge curve projection approach**

- The results reveal
  → **FCN-only labels significantly outperform the original and FCN+Edge label data**
  → **The new labels significantly enhance accuracy and reliability, providing improved vocal-tract label data for the research community**

- The refinement labels data proposed in this study available at
  → **https://github.com/ha3ci-lab/USC-TIMIT_rtMRI_Landmarks**
  → **It can be utilized as auxiliary information for the current existing USC-TIMIT dataset**

Thank you


JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990


Nara Institute of Science and Technology
NAIST