Tunghai International Conference on Second Language Teaching and Research (SLTR.THU)
Oct. 21, 2023, Taichung, Taiwan

Ultrasonic-and-optical-imaging-assisted automated speech correctness judgment model

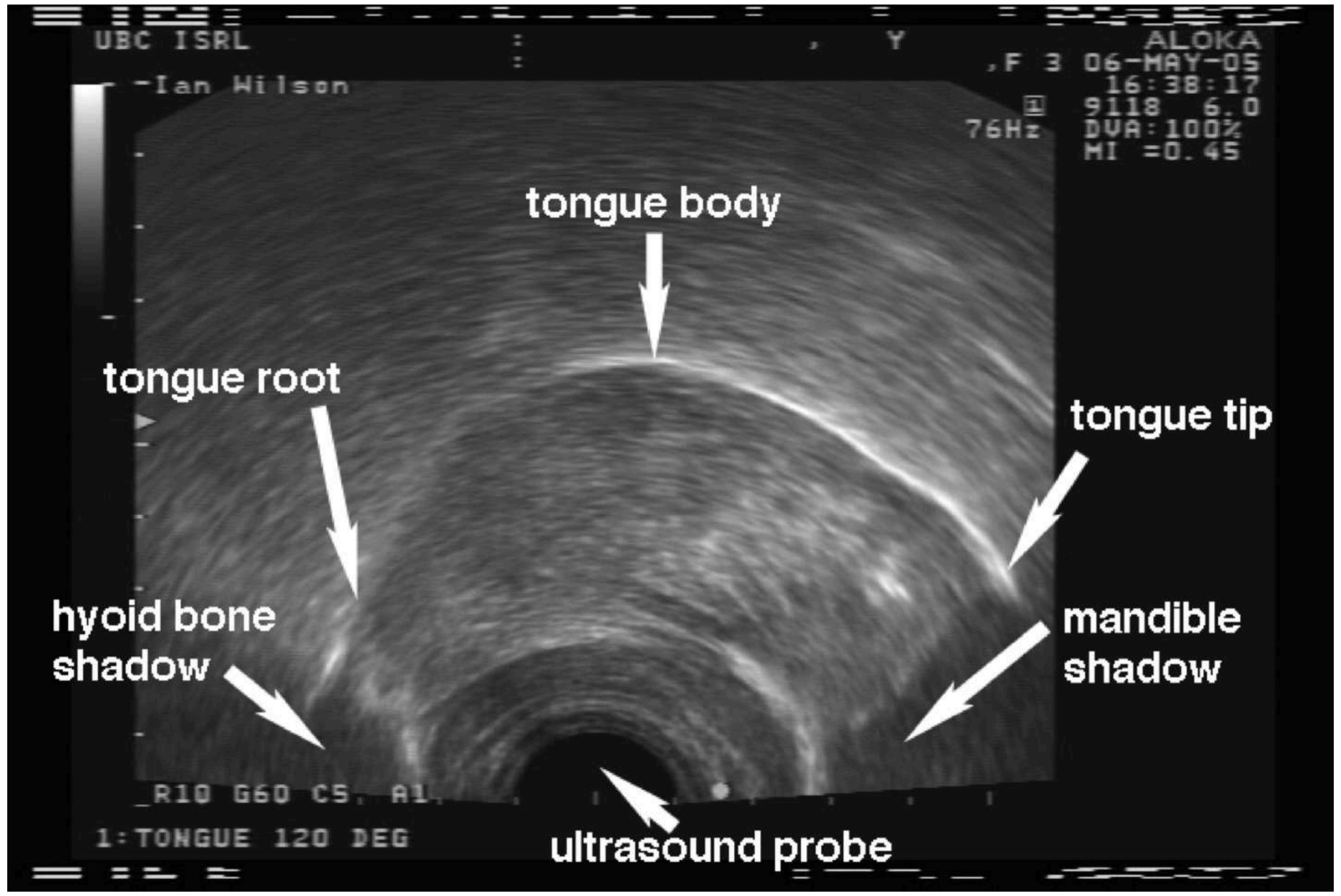
Huang, Po-Hsuan 黃柏瑄[1], Liu, Yi-Chen 劉倚辰[2]

Graduate Institute of Linguistics 語言學研究所[1]
Department of Chemical Engineering 化學工程學系[2]
National Taiwan University 國立台灣大學

Introduction

The use of ultrasound in second language acquisition

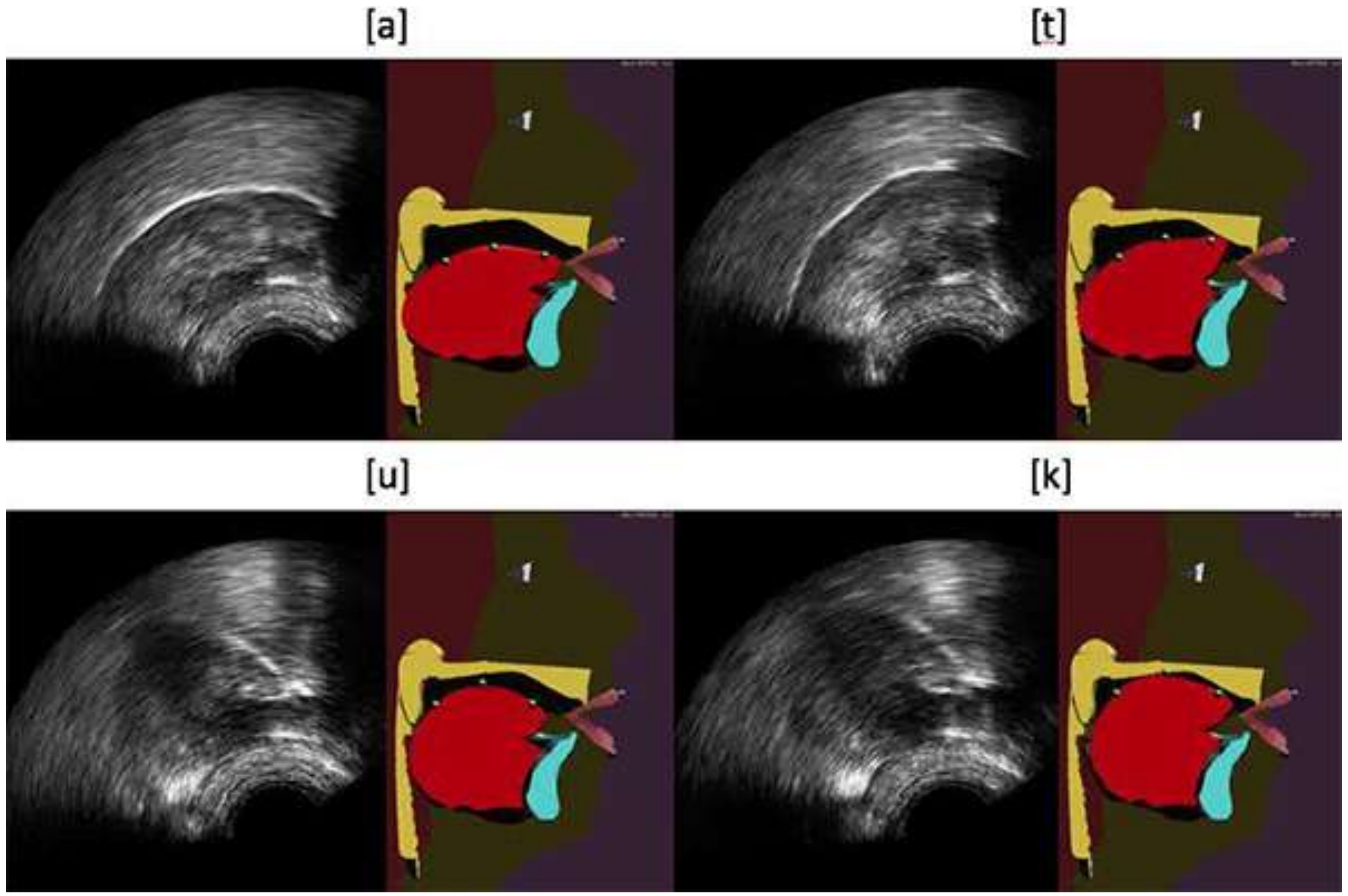
- Ultrasound (US) :
  - ultra-high frequency sound waves
  - can be used to visualize tongue surface



Introduction

The use of ultrasound in second language acquisition

- Different sounds tend to have distinctive postures.



Introduction

The use of ultrasound in second language acquisition

- US has been used to help L2ers approach correct articulatory gestures by visualizing the tongue contours(Gick et al. 2008).

- A typical process of ultrasound-assisted sound training:

Pre-training recording → Intervention → Post-training recording

↓

Evaluation

Introduction

The use of ultrasound in second language acquisition

- US has been used to help L2ers approach correct articulatory gestures by visualizing the tongue contours (Gick et al. 2008).
- A typical process of ultrasound-assisted sound training:

Pre-training recording → Intervention → Post-training recording

↓

Evaluation

Introduction

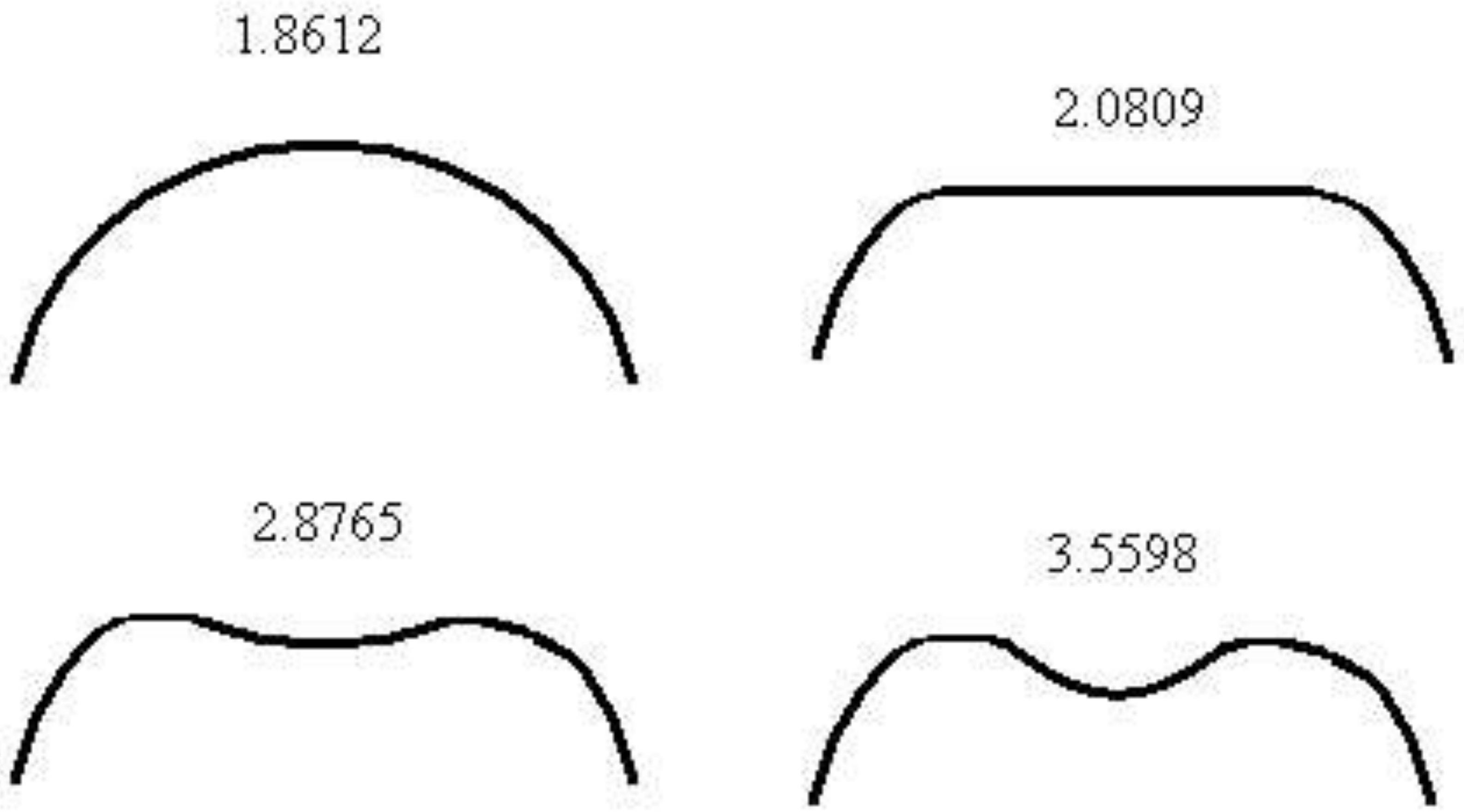
The use of ultrasound in second language acquisition

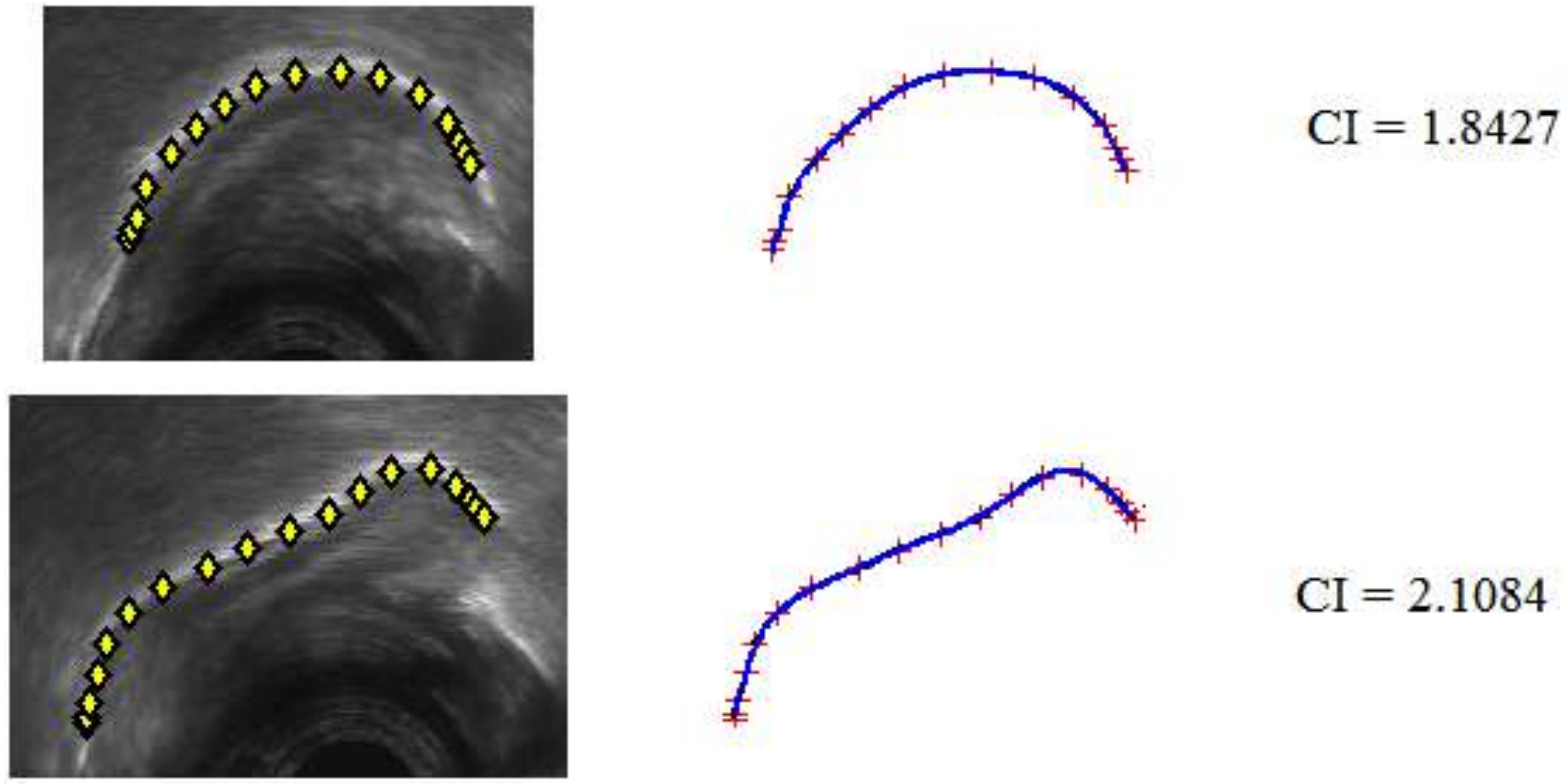
- Usually during the intervention and evaluation:
  - a fixed set of criteria has to be made to evaluate the production.
  - instructor(s) are usually required to assist the L2er in judging the goodness of their production based on these criteria.
  - these criteria may also need to be calculated with traced tongue contours.

Introduction

The use of ultrasound in second language acquisition

- For example, the correctness of /r/ may be assessed by calculating the curvature index (CI) of the tongue.





Stolar & Gick (2013)

Introduction

The use of ultrasound in second language acquisition

- As a consequence, the traditional usage of US:
  - Qualitative measure:
    - requires trained professionals.
    - may exist individual differences for different instructors.
    - requires post-training evaluation from instructors/naïve native speakers.

Introduction

The use of ultrasound in second language acquisition

- As a consequence, the traditional usage of US:
  - Quantitative measure:
    - requires post-training tongue tracing and calculation.

Introduction

The use of ultrasound in second language acquisition

- Importantly, the evaluation:
  - requires a predetermined set of criteria.
  - cannot be done in real time.

Introduction

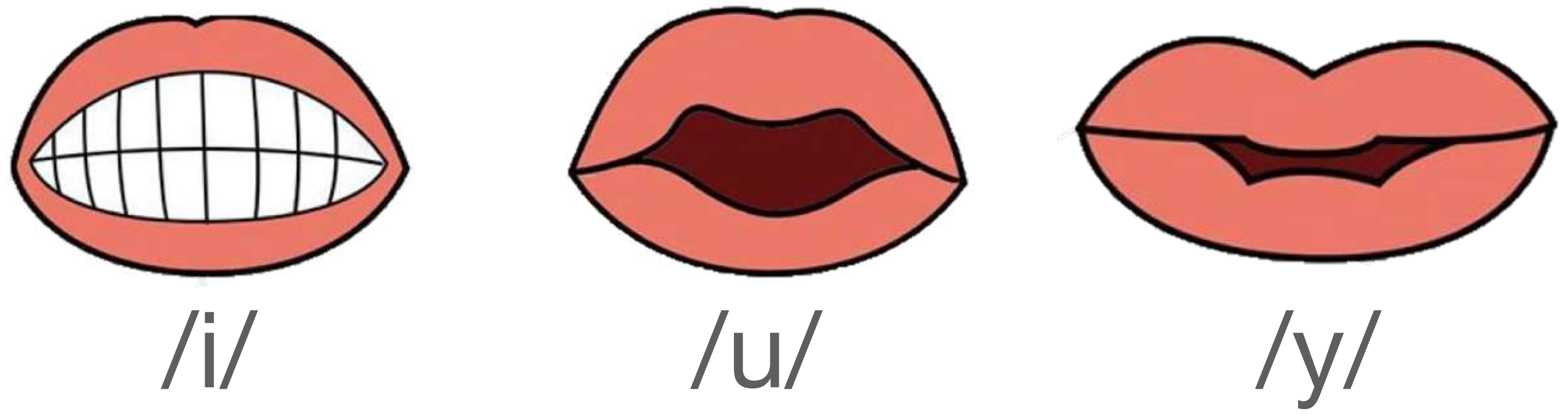
The potentials of neural networks

- An automated neural network could:
  - make consistent evaluations.
  - provide the goodness of production in real time.
  - avoid the need for a trained professional.

Introduction

Ultrasound-only training vs. combination w/ optical input

- A more obvious downside of US is that it lacks the information of lips.
- Several languages distinguish between rounded/unrounded sounds:
  - Vowels: e.g., Mandarin /i/ vs. /y/
  - Consonants: secondary coarticulation (e.g. X^w)



National Taiwan University 國立臺灣大學

Introduction

Ultrasound-only training vs. combination w/ optical input

- Combination w/ optical imaging can provide more thorough assessment.

Goal

Automated assessment w/ ultrasound + optical imaging

	Traditional US-assisted training		Automated assessment w/ US-optical imaging
	Qualitative	Quantitative	
Requirement of fixed criteria	Yes	Yes	No ✓
Requirement of instructors during intervention	Yes	Yes	No ✓
Pos-hoc assessment	Evaluation from instructors/ naïve native speakers	Tongue-tracing+calculation	Not required ✓
Individual difference	Yes	No ✓	No ✓
Timeliness	May be quick w/ trained experts ✓	Slow	Real-time ✓
Lip information	No	No	Yes ✓

Methods

- **Material selection**
  - Sound type selection
  - Number of repetitions
- **Data collection**
  - Ultrasound recording
  - Camera recording
  - Audio recording for alignment
- **Data preprocessing**
  - Alignment of US & optical imaging
  - Segmentation
  - US image preprocessing
  - Optical image preprocessing
- **Model building**
  - Model structure design
  - Training
  - Fine-tuning

Methods
Material selection

- Vowels were chosen to be test cases：
  - relative invariance across time.
  - quantifiable continuous properties：tongue height, tongue frontness, lip roundedness

Methods

Material selection

- One naïve native speaker of Mandarin (male, 24) and one trained phonetician (male, 26) were recruited to produce the data for model training.
- All vowels on the IPA vowel chart were produced by the trained phonetician. Mandarin vowels were produced by the naïve Mandarin speaker.
- Repetition: 10 times per vowel for 10 seconds.

Methods

Data collection

- Apparatus:
  - Ultrasound:
    - CGM OPUS 5100
    - 37 fps
    - Reception frequency: 4-8.3 MHz
  - Camera:
    - iPad Pro
    - 120 fps
    - Resolution: 1080
  - Audio:
    - USBPre2
    - Sampling rate: 44100
    - Saved as .wav

Methods

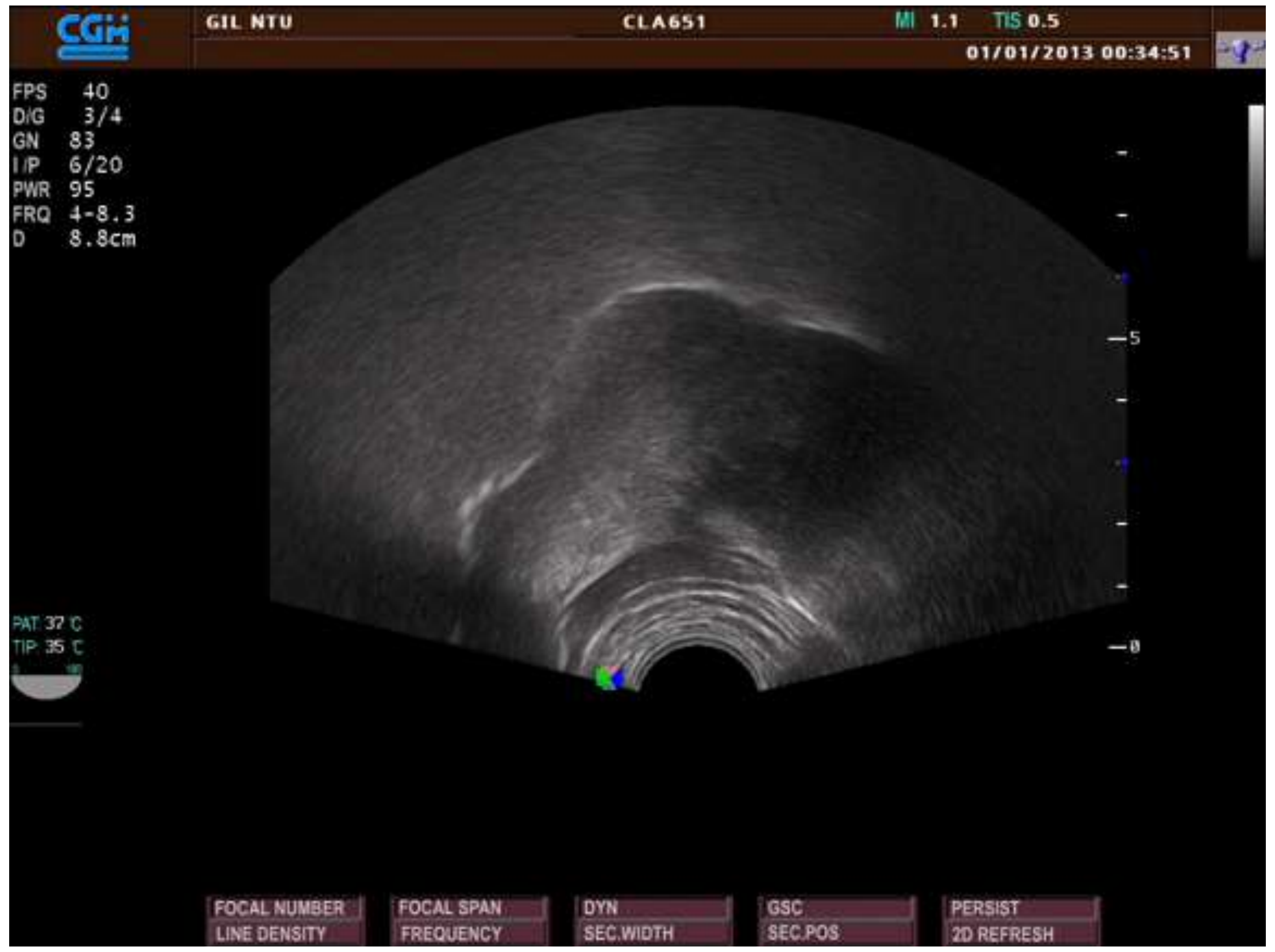
Data preprocessing

- Video alignment based on audio.
- Vowel segment onsets/offsets were marked with Praat's TextGrid.
- Image extraction:
  - All ultrasound/optical frames within the vowel segments were extracted.

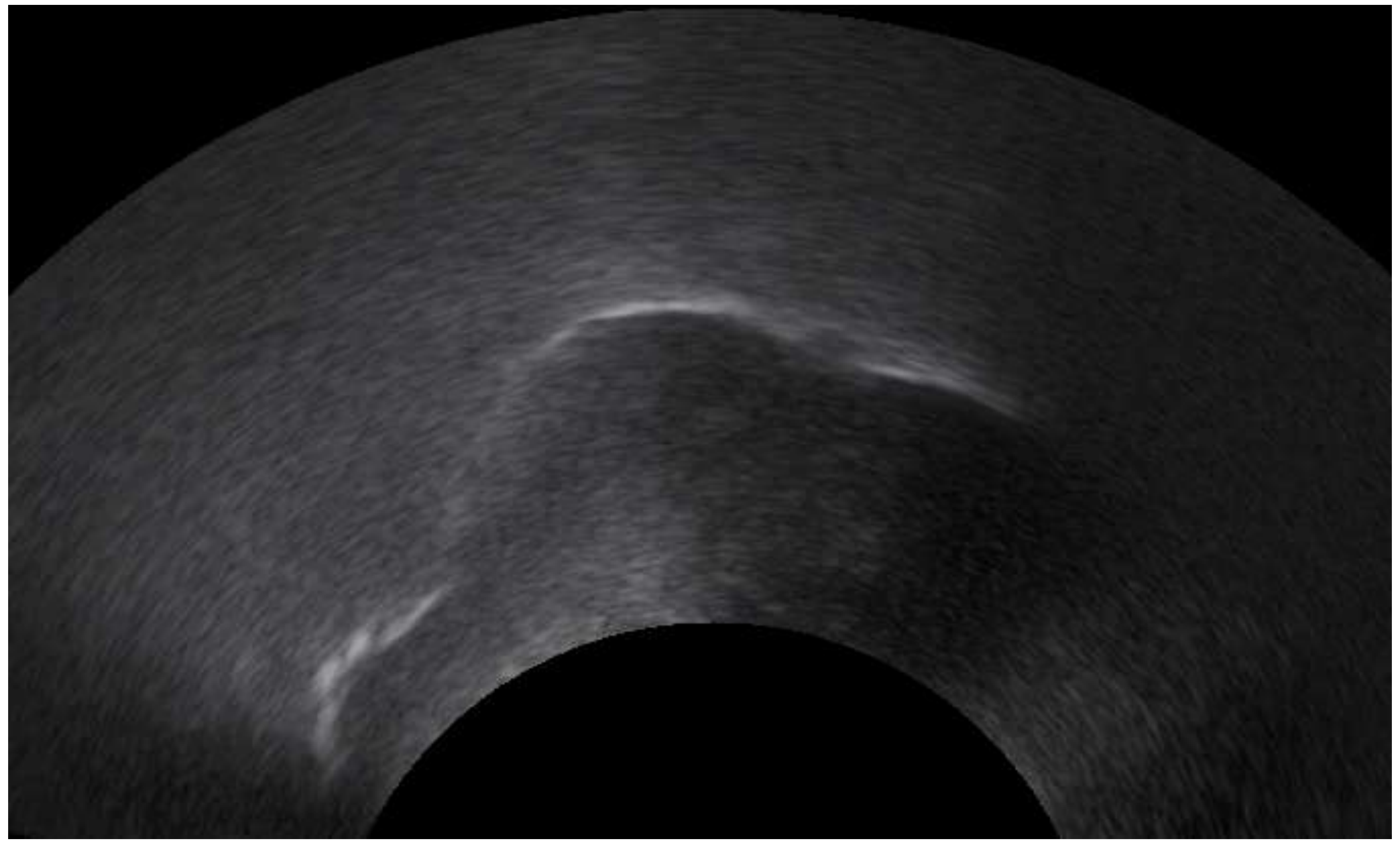
Methods

Data preprocessing

- Ultrasound frames were masked to focus on the region of interest.



Original Ultrasound Frame

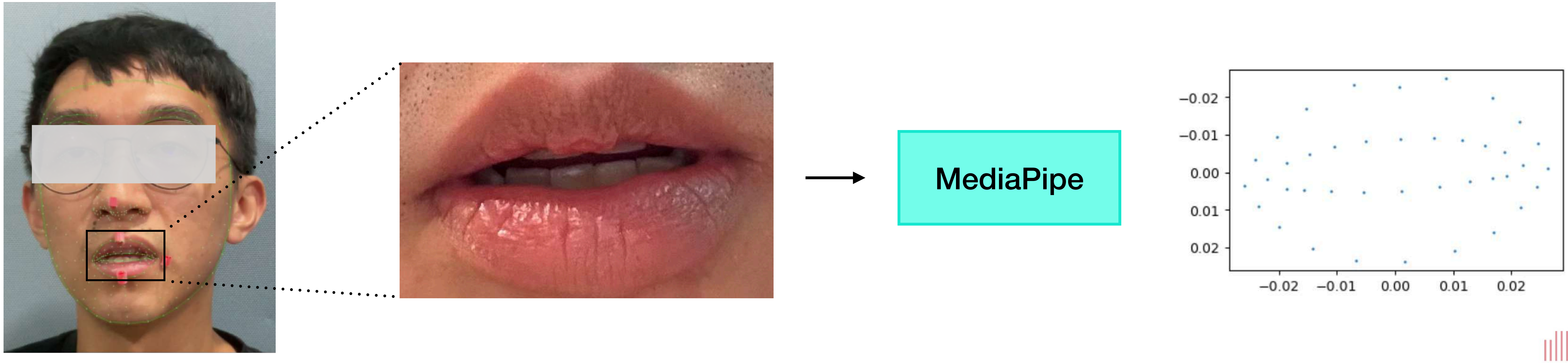


Preprocessed Image

Methods

Data preprocessing

- Lip information was extracted with 42 landmarks marked w/ MediaPipe.



Methods

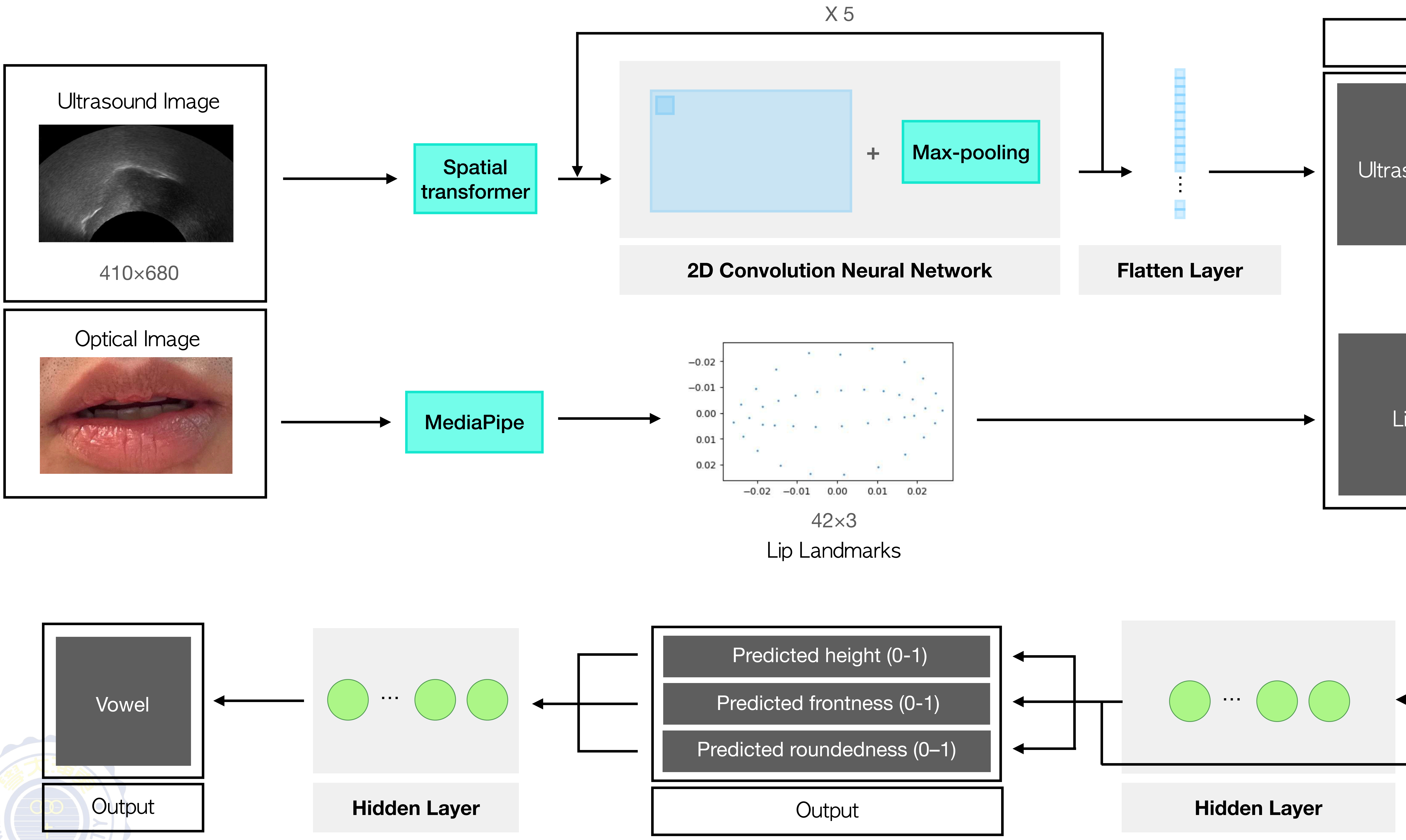
Model building

- A spatial transformer network and 2D CNN were used to deal with ultrasound images.
- Flattened feature maps and traced lip landmarks were then used as input.

Methods

Model building

- The model predicted four metrics:
  - tongue height (float, 0–1)
  - tongue frontness (float, 0–1)
  - lip roundedness (float, 0–1)
  - target vowel (categorical one-hot encoded array)
- Evaluation metrics:
  - Mean squared error (MSE)
  - Accuracy


X 5
Ultrasound Image
410×680
Optical Image
Spatial transformer
MediaPipe
+
Max-pooling
2D Convolution Neural Network
Flatten Layer
−0.02
−0.01
0.00
0.01
0.02
42×3
Lip Landmarks
Input
Ultrasound Data
+
Lip Data
Vowel
Output
Hidden Layer
Predicted height (0-1)
Predicted frontness (0-1)
Predicted roundedness (0–1)
Output
…
Hidden Layer

Results


predicted vowel: /u/
height: 0.997
frontness: 2.889
roundedness: 0.999
GIL NTU
CLA651
MI 1.1
TIS 0.5
01/01/2013 00:35:37
FPS 40
D/G 3/4
GN 83
I/P 6/20
PWR 95
FRQ 4-8.3
D 8.8cm
PAT: 37 ℃
TIP: 35 ℃
FOCAL NUMBER
LINE DENSITY
FOCAL SPAN
FREQUENCY
DYN
SEC.WIDTH
GSC
SEC.POS
PERSIST
2D REFRESH

Results

Model evaluation

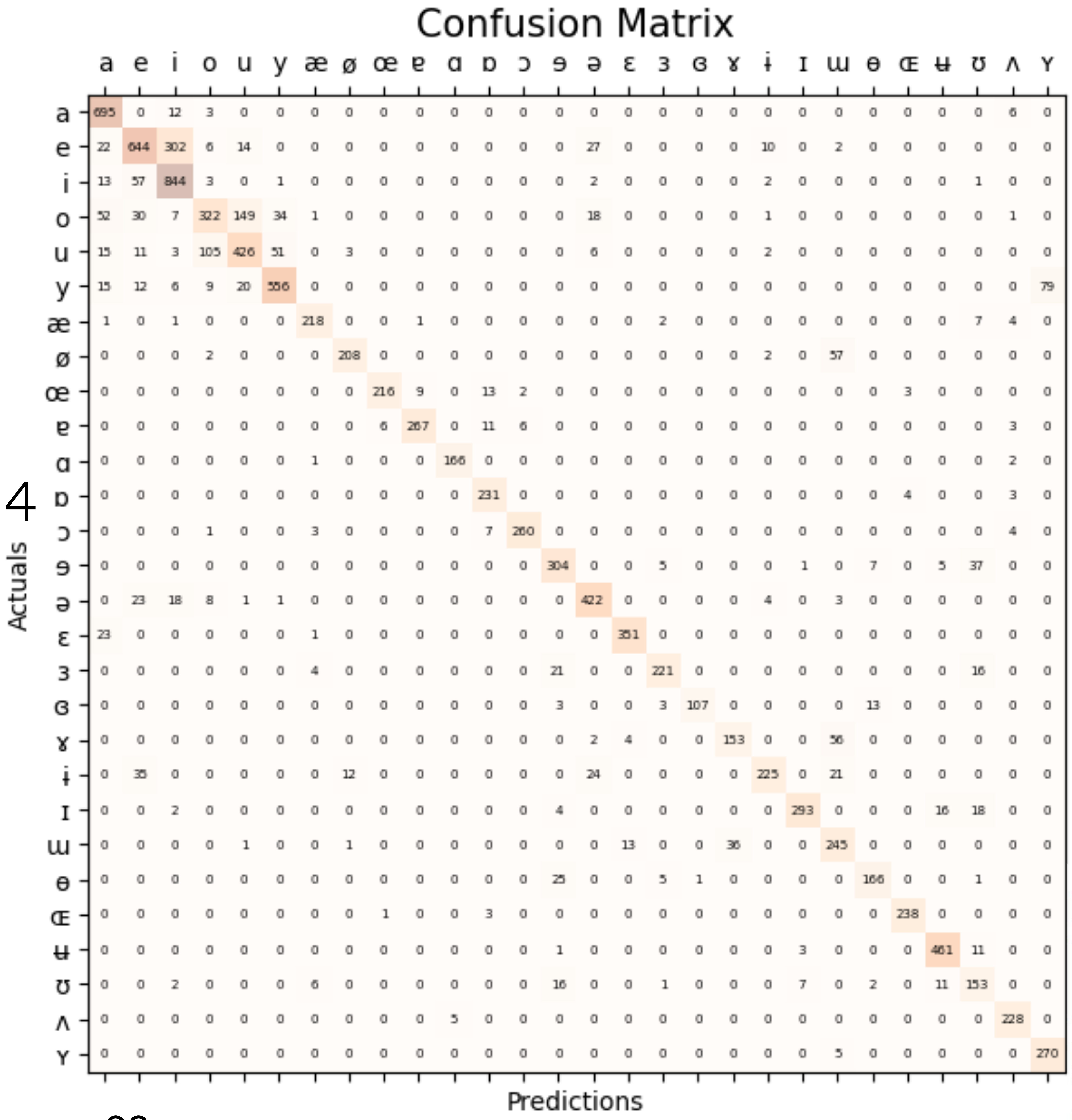
- Vowel prediction evaluation
  - Accuracy: 0.830
  - F1 ($\frac{2 \times precision \times recall}{precision + recall}$): 0.854
  - Recall ($\frac{TP}{TP + FN}$): 0.855
  - Precision ($\frac{TP}{TP + FP}$): 0.862

		Reality	
		X	Not X
Prediction	X	True positive (TP)	False positive (FP)
	Not X	False negative (FN)	True negative (TN)

Results

Model evaluation

- Vowel prediction evaluation
  - F1 ($\frac{2 \times precision \times recall}{precision + recall}$) : 0.854
  - Recall ($\frac{TP}{TP + FN}$): 0.855
  - Precision ($\frac{TP}{TP + FP}$): 0.862
- Accuracy: 0.830



Results

Model evaluation

	Tongue frontness	Tongue height	Lip roundedness
MSE	0.031	0.015	0.031
MAE	0.077	0.068	0.064
R^2	0.813	0.882	0.871

Discussion

- The results show the potential for automated speech correction systems.
- Combination of ultrasound and optical imagining promotes a more complete assessment of the goodness of production.

Further research

- Larger dataset
- More diverse participants
- Vowel production from native speakers

More about us

Biomedical and Tissue Engineering Laboratory
生醫與組織工程實驗室

Speech Behavior and Science Lab
語音行為與科學實驗室

National Taiwan University
國立臺灣大學